万卷方法

SURVEY RESEARCH METHODS

调查研究方法

■（美）弗洛德·J·福勒,Jr 著
■孙振东 龙藜 陈荟 译

第3版

重庆大学出版社

责任编辑：雷少波
封面设计：黄　河

本书将在以下方面为您正在或即将开展的调查工作提供帮助：

- 尽可能提高您调查的回应率，无论您采用问卷调查，还是当面访谈，或者是电话调查、网络调查。
- 尽可能提高您在设计题项上的信度和效度，这是保证您的调查取得成功的关键。
- 如果您需要使用访谈员协助调查的话，那么本书关于访谈员工作的有关章节必将对您的调查工作有所助益。
- 本书还在调查资料的整理、调查的评价等方面提供可行性建议。



ISBN 7-5624-3289-9
定 价：15.00元



SURVEY RESEARCH METHODS

调查研究方法

■（美）弗洛德·J·福勒,Jr 著
■孙振东 龙藜 陈荟 译

第3版

重庆大学出版社

图书在版编目(CIP)数据

调查研究方法/(美)福勒(Fowler, F. J)著；
孙振东，龙藜，陈荟译. —— 重庆：重庆大学出版社，2004.10
(万卷方法)
ISBN 7-5624-3289-9

Ⅰ.调… Ⅱ.①孙…②龙…③陈… Ⅲ.调查研究—方法
Ⅳ.C31

中国版本图书馆 CIP 数据核字(2004)第 106115 号

调查研究方法
(第 3 版)
弗洛德·J·福勒，Jr 著
孙振东 龙 藜 陈 荟 译
责任编辑：雷少波 版式设计：周 晓
责任校对：邹 忌 责任印制：秦 梅

*

重庆大学出版社出版发行
出版人：张鸽盛
社址：重庆市沙坪坝正街 174 号重庆大学(A 区)内
邮编：400030
电话：(023) 65102378 65105781
传真：(023) 65103686 65105565
网址：http://www.cqup.com.cn
邮箱：fxk@cqup.com.cn (市场营销部)
全国新华书店经销
重庆科情印务有限公司印刷

*

开本：890×1240 1/32 印张：5.875 字数：162 千 插页：2
2004 年 10 月第 1 版 2004 年 10 月第 1 次印刷
印数：1—4 000
ISBN 7-5624-3289-9 定价：15.00 元

作译者简介

弗洛德 J·福勒,Jr 毕业于 Wesleyan 大学，于 1966 年获得密歇根大学社会心理学博士学位。此后,他大多数时间都是在马萨诸塞一波士顿大学的调查研究中心度过的。他在这个中心参加了许多调查项目的设计和施行,这些项目的主题极其广泛,包括:对地方政府和机构的态度、赌博、种族冲突、犯罪的恐惧、犹太人的识别、冒险的理解以及从健康计划中获得医疗服务的经历等。除此之外,还涉及其他更多的研究主题。最近他的方法研究主要集中在以下三个领域:调查问题的设计和评估、如何减少调查资料里与访谈员有关的误差和如何测量医疗服务对病人的影响。他在哈佛的公共健康学院以及其他地方教授调查研究方法。他担任调查研究中心的主任,长达 14 年之久。除了已发表的大量研究论文外,他还是《标准化调查访谈》(*Standardized Survey Interviewing*, 与 T.W. Mangione 合作)和《改进调查问题》(*Improving Survey Questions*)两本被广泛使用的有关调查研究方法的专著的作者。

孙振东 教育学博士,西南师范大学教育科学学院副教授。主要研究领域为教育基本理论,教育研究方法与方法论。

龙 藜 西南师范大学教育学原理专业硕士研究生，研究方向为教育基本理论。

陈 荟 西南师范大学课程与教学论专业硕士研究生，研究方向为比较教学论。

为方法理性鼓与呼*

——万卷方法策划报告暨出版说明

究竟是中国内地学界对于研究方法的漠视，导致研究方法出版物的匮乏？
还是研究方法出版物的匮乏，导致学界没有对方法问题予以应有的重视？
这是一个鸡生蛋蛋生鸡的问题。
作为图书出版的编辑人和策划人，对之多论无益。
但，作为图书出版的编辑人和策划人，我们却非常乐意——为方法理性鼓与呼！

我们乐于，也更善于从出版物的数量和质量的比较中，来了解和表述某一类知识的生产和传播情况，以此作为我们图书出版策划的基础。同样，在万卷方法的策划之前，我们便对美国、中国台湾及中国内地三者，1999—2002年间关于社会科学研究方法的图书出版情况做了一个简单的比较：

美国在社会科学研究方法方面出版物的数量之多，至少足以让我们惊讶！由于不便统计，在此，仅就美国的SAGE出版公司在社会科学研究方法方面的图书出版情况，做一个简单介绍，以窥其一斑。SAGE出版公司4年中出版的社会科学研究方法方面的书籍（包括再版书）便有两百余种，其中，既有一两百万字的大部头工具书，如Handbook of Qualitative Research，Handbook of Research Design & Measurement；也有10万字左右的口袋书，如仅一个“应用社会科学研究方法系列丛书”（Applied Social Research Methods

* 本文原载《中华读书报》2004年3月31日，第15版。在此，根据需要作者对原文做了少量改动。

Series)便有49个品种，其中每本书对于案例研究方法、调查研究方法、网络调查方法等做了分门别类的介绍。其覆盖面之广，研究之深入、具体、操作性强是我们所难以想象的。如果说，我们与美国的差距尚可理解和接受的话，那么，同一时期，中国台湾在这一方面的研究和传播情况也足以令我们汗颜。4年间中国台湾出版了大量深入研究社会科学研究方法方面的书，其中不但有本土作者的原创作品，也有从英美等西方国家引进版权的相对比较成熟的社会科学研究方法方面的作品。更为可贵的是，其中有几家出版公司已经注意从社会科学研究方法的体系着眼来组织自己的策划和出版，在这方面的出版有了一定的规模，内容的组织也显得比较成熟：如韦伯文化事业出版社的"研究方法系列丛书"、弘智文化事业公司的"应用性社会科学调查研究方法系列丛书"等都是发展比较成熟、规模比较大的专门传播社会科学研究方法的系列丛书。

而中国内地方面，同期虽然也出版了一批研究方法的书籍，但无论是规模还是研究内容的深入丰富程度都无法与美国及中国台湾两地相提并论。从规模看，这期间中国内地出版的社会科学研究方法类的著作也就四五十本，难望美国同行之项背，仅与中国台湾同期出版规模大致相当。从此类出版物内容的深入丰富程度来看，大陆出版的社会科学研究方法类出版物主要集中在各学科内部，如教育研究方法、心理学研究方法、社会学研究方法、经济学研究方法、体育科学研究方法等，每本书都面面俱到谈调查、谈案例、谈访谈、谈田野、谈观察等，而很少有对某一种方法进行深入研究

的图书，更没有像美国或中国台湾那样形成比较系统的研究社会科学研究方法的系列丛书。在这一领域，这样的图书结构对于应付大学本科生在研究方法方面的入门性需求(即作为教材)还行，但对于迅速培养一大批拥有科学、理性的研究头脑的学者，特别是对于青年学者，如硕士、博士研究生的成长则是远远不够的。其致命的弱点表现在三个方面：其一，一本书铺得太开而无法细化深入，以至于绝大多数学生虽然上了方法的课，却除了记住了几个名词和对一些方法的理论及应用略知一二之外，至于什么情况下选择什么方法最恰当、具体怎么操作、这种方法可能会有哪方面的不足需要加以处理等研究细节问题仍然处于无理性状态。其二，既然我们的方法建设囿于学科内部，而不能对方法进行纵深的开发，那么，反正学科就那么几个，于是方法书大多是低水平重复。比如教育研究方法的书，现在全国使用做教材的至少也有十多本吧，但你若有心思看的话，一本足矣！其三，出版界对研究方法图书这种淡淡的一笔带过的做法，不仅很难形成一种风气，从意识深处提升起大家对于研究方法的关注，而且更加剧了整个学界对于研究方法的漠视，许多学者只有在著书立说之时才想到似乎从“作品结构的完整性”上看应该谈及方法问题。

社会科学界近年流行两个词语：一曰反思，二曰接轨。所谓反思也即是对于学科的理论基础、学科的结构、学科的历史与未来等问题的全面梳理；所谓接轨也即是希望中国的社会科学研究能够融入世界社科研究的大潮中，与国际上的社会科学研究实现彻底

的、清晰的对话。在此，我们以为，无论是反思还是接轨，中国的社会科学界和传播界都必须投入一些精力来关注、研究、教授和传播社会科学研究方法。惟如此，才能在中国形成完善、科学的社会科学研究方法体系和学科群，也才能使对研究方法关注与理性应用在中国社会科学界深入人心、蔚然成风；惟如此，也才能为中国社会科学研究与国际接轨搭建一个平台。

以上种种，促使我们策划出版了万卷方法系列丛书，献给忠实于人文社会科学研究的人们！

雷少波　　崔祝

2004 年 6 月于重庆大学出版社

英文版前言

本书这个版本的出版目的与其前两个版本一样，都是为那些原本不是统计学者或方法论学者的人们提供一个有关调查中误差来源的一般知识的全面总结。我认为，那些重要的知识，诸如良好的样本、高回应率、好的问题设计、良好的访谈过程、有效的资料收集方案，以及它们对调查资料质量的影响等，从本书的第 1 版发行至今 16 年来并没有太多变化。弄清这些问题的重要性以及它们如何影响了调查估计质量，对于开展调查或者使用他人收集的调查资料而言都是非常重要的。对此，相信读者在读完本书后就会明白。

我做了相当大的努力以使本书对一般的读者来说都能清晰易懂。对社会科学研究的了解会帮助读者理解本书，但并非一定要有相关知识的人才能读懂本书。

本书是简写本，必然地，我就要对文章内容的详细程度做出选择。如果有读者希望对特定的问题诸如抽样、调查手段设计或者资料收集方案等有更深入的了解，那就需要进一步参看一些大部头的专著。

这一版本与前面版本相比，主要有三方面的不同：第一，本书包括了在最近 8 年里发表的方法论资料。第二，电脑的革新和互联网的发展给研究者的选择方式以及问题的处理带来了改变。这些内容有必要增添进来，以使本书跟上最新的科研动态。第三，评估调查问题的技术仍在不断发展，这是调查方法不断得到改进的一个领域。新版本体现了这些方面的进展。

在以前的版本里，我对罗伯特•凯恩（Robert Kahn）、莫里斯•

阿克塞尔罗德（Morris Axelrod）以及后来的安格斯•坎贝尔（Angus Campbell）所给予的帮助表达了我诚挚的谢意。新版本又多处参考了罗伯特•格雷夫斯 （Robert Groves）、西摩•萨德曼（Seymour Sudman）、诺曼（Norman）、布拉德伯恩（Bradburn）和唐•迪尔曼（Don Dillman）的著作。读了本书之后，想更多地了解调查研究方法的学生应当再去阅读上述人们的专著，这些专著都对方法论知识做出了很大的贡献。最后，我成长为一个方法论学者深受与查尔斯•F•坎内尔（Charles F.Cannell）共事的影响，他对我们理解访谈员和调查手段设计的贡献极其巨大，他的工作对我的调查研究方法观点的影响是深远的。

关于本书，我要感谢玛丽•艾伦•科尔顿（Mary Ellen Colton）、汤尼•罗曼（Tony Roman）、维基•斯特林费洛（Vickie Stringfellow）和加拉赫•加拉格尔（Trish Gallagher）对本书的仔细审阅并对很多章节做出的有益的评点；感谢调查研究中心为我完成本修订版提供的时间和援助；感谢朱蒂•钱布利斯（Judy Chambliss）一直以来给予我的帮助，她使我在写作过程中得以保持健康的精神状态。尽管人们给予我巨大、有益的帮助，但本书的最终责任应由我个人承担。

杰克•福勒（Jack Flower）

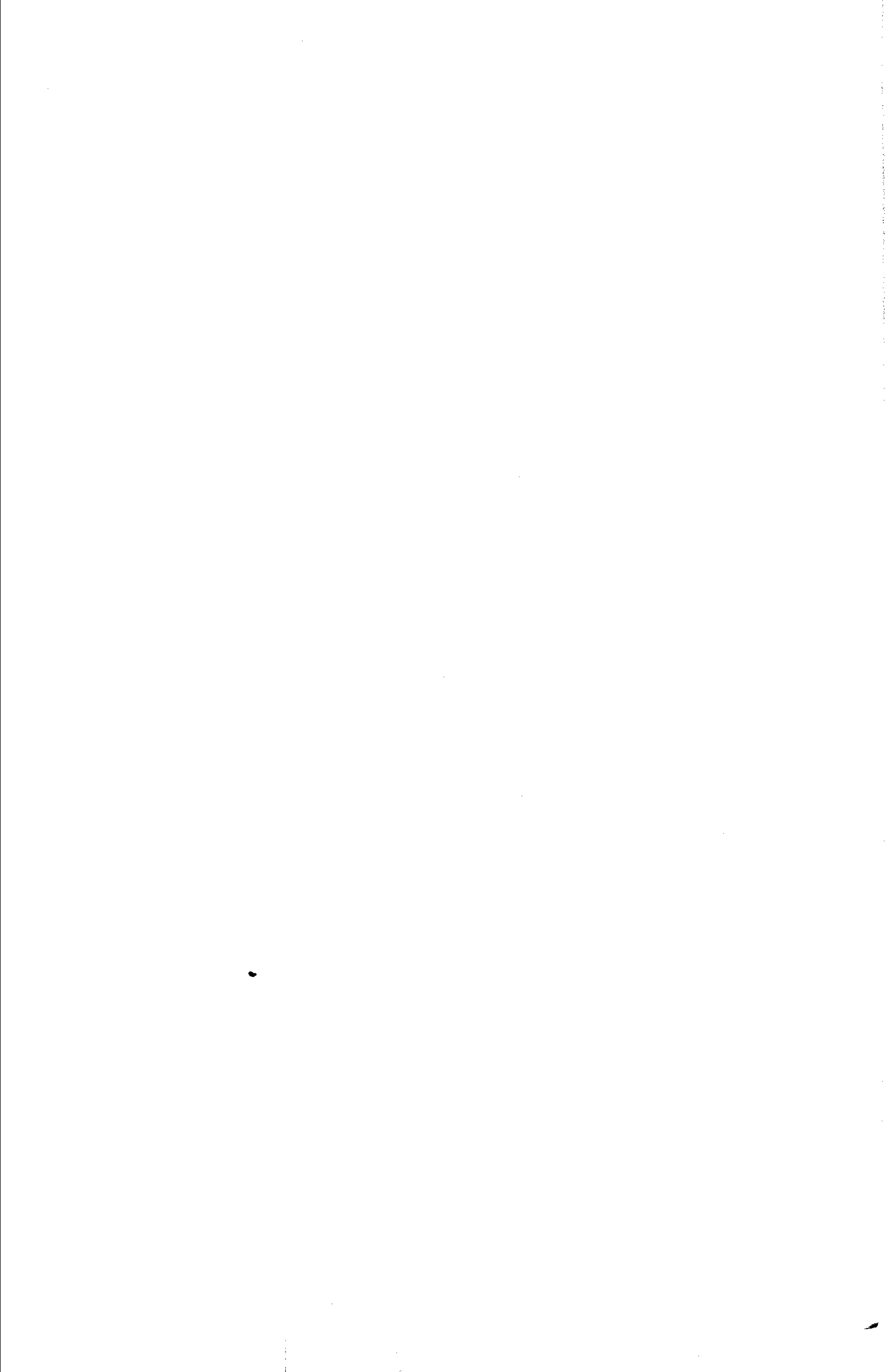
目　录



1 概 述

Introduction

本书讨论了对一个人群样本提问以获得统计性描述资料的调查的规范和实施程序。调查包括抽样、问题设计以及资料收集方法。那些想收集、分析或阅读调查资料的读者将从中认识到调查的每一个方面的细节是如何影响整个调查的精确度、准确度和可信度的。



本书的主题是各种社会调查中的资料收集，它包括调查的一般程序、良好的实施规范以及各种保证调查资料质量的设计决策的运用。本书的目的在于为那些想收集、分析和阅读调查资料的人们提供评估资料收集过程的可靠基础。读者将逐步了解到资料收集中的细节与调查数据和统计资料的可信度之间关系的各个方面。

许多资料收集和测量过程都可称作调查，但本书关注的调查具有以下特征：

- 调查的目的是产生统计数据，也就是说用定量的或数字的描述方式来体现研究总体的某些方面。
- 收集资料的主要方式是向人们提问，回应者的答案组成了待分析的资料。
- 一般说来，收集的资料只是来自总体的一部分成员，即来自一个样本，而不是来自总体中的每个成员。

调查原因

美国宪法明确指出，每十年必须进行一次符合所列举标准的人口普查。在每十年一次普查中，关于人口状况的统计资料都是通过向人们提问获得的，而不是采取抽样方式，调查资料被认为是从总体中的每个人那里收集到的。

十年普查的目的是计算人数，以之作为确保众议院中有合适的代表人数的基础。然而，普查也已经变成用于其他目的的一个主要资料来源。除了简单的记数统计，它也收集调查对象的种族、年龄、家庭组成、教育情况、家庭类型以及调查对象所具有的其他特征的资料。

十年普查的内容已被大大丰富了，其目的是为了满足政府部门和研究员对描述性资料的需要。但是，它只包括了人们想知道的关于总体的一小部分资料，并且由于每十年才进行一次，其价值非常有限。为了弥补这些资料上的空白，从 20 世纪 30 年代起，专门目的的调查在美国社会中流行起来。

大多数人熟悉这三种调查技术的使用：调查公众对报纸、杂志上刊登文章的意见；调查公众的政治观念和思想以帮助候选人竞选；进行市场调查以弄清消费者的偏好和兴趣。每一个充分准备的调查研究项目的最主要目的是获取公众的主观感受。另外，还有很多关于人们行为和情况的事实就只有通过让调查对象自己谈论而获得。调查研究方法几乎适用于所有的公共政策。以下简略列出了一些主要应用方面：

- 失业率是由劳工统计局以及许多其他收集关于职位和工作的统计数据的机构例行发布的资料，它是基于人口普查局进行的家庭调查获得的。类似的商业和工业调查的目的是描述产品和人力资源的需求。
- 通过调查才能获得可靠资料的另一个领域是人们的收入和支出方式。消费者的消费方式及需求已被证明为经济发展的重要预示。
- 自 20 世纪 50 年代后期以后，人口普查局公共健康服务中心施行了国家健康访谈调查，这项调查收集了有关健康状况、医疗服务的使用以及危及健康的行为的基本资料。这些都是只有通过可靠的调查研究才能获得满意资料的课题。
- 犯罪事件的主要资料的来源通常是警察局的记录，然而这些记录仅包括了报案的案件。对于大多涉及受害者的犯罪，其犯罪率和受害者性格的更可信的测定，是通过调查实现的。在 20 世纪 70 年代，为提供此方面的数据，实施了全国犯罪行为调查。另外，调查是测量人们对犯罪的关注和惧怕情况的惟一方法。
- 由美国农业部进行的调查是众多历史悠久的调查应用之一。农业部调查农民以评估不同农作物的种植率、预测不同食品的产量。

精神健康状况、交通的需要和使用类型、政治行为、家庭特征，这些都是调查研究可以广泛推行的一些例子。毫无疑问，在美国，最大的调查资料收集者就是联邦政府，特别是人口普查局和农业部。另外，每年还有成千上万的个人调查在进行。它们是由大学、非赢利或赢利的调查机构组织的。

为特殊目的而进行的调查资料的收集是一种特别昂贵的解决资料缺乏问题的方法。在进行有关工作之前,需要十分彻底地查找其他资源中是否已经存在所需要的资料。虽然有一些人认为调查是了解总体的第一项努力,然而全面的概率样本调查只能在确定没有其他办法可收集资料的情况下才进行。即使采用了这种保守的收集方法,也常发现只有特殊目的调查才能提供所需的资料。除了能获得以其他方式不能获得的资料以外,规范的调查得到的资料还有优于从其他来源所获资料的三个特点:

- 概率抽样(probability sampling)使人能确信样本没有偏误,并能评估资料的精确度。从通过适当方法选定的样本里获得的资料,比从诸如参加会议的人群、呼声最大者、来信者或者碰巧顺便投票的人等构成的样本里获得的资料更真实。
- 统一的标准化测量(standardized measurement)可以保证从调查对象那里收集到可比较的资料。没有这种标准化测量,是不会产生有意义的统计数据的。
- 特殊目的的调查也许是确保获得一个给定的分析需要的所有相关资料的惟一途径。即使关于某些事实已经有了资料,但是它们可能适合分析所需的其他特征。例如,医院的出院结账记录不会有病人经济收入的记录,因此,为了研究人们的收入与住院经历之间的关系,就需要收集人们的收入与住院情况的调查。

通常对于已确定的主题都可以从人们的谈话、印象或记录中得到现成的资料,这些资料也常存在一些不足。除了对资料需求的评估外,决定进行调查的因素还包括考虑工作人员的来源问题。如果所需的工作人员和专家不足或聘请不到,那么最后调查所得的资料就不会太准确。这就将引出了下一部分探讨的内容:一个优秀的调查是由哪些部分组成的。

调查的构成

正如所有科学测量一样，社会调查测量不会是没有误差的。调查过程中的每一个步骤，最终都会对资料准确描述对象的可能性有重要影响。

一个样本调查包括抽样、问题设计和资料收集三套不同的方法。这些活动的每一项除了都适用于样本调查外，还有许多其他的用途，但是它们的结合对良好的调查设计起着举足轻重的作用。

抽 样

普查是收集总体所有成员的资料。使调查有实用价值的一大进步是人们知道了如何抽样——从总体中选择一小部分代表整个总体。良好抽样的关键就是找到一种方法让总体中所有（或几乎所有）的成员都有同样（或已知）的几率被抽样，即使用概率方法选择样本。早期的调查和民意测验通常是使用便于得到的样本或是忽视了总体中的重要部分，这些方式都不能提供可靠、可信的数据。

农业部切实发展了能够提供特定区域内居住人口的可靠性数据描述的全面的概率抽样程序。这一程序是在为预测农作物产量而设计抽取土地样本的工作中发展起来的，抽取一些家庭单位或其中的居民作为样本只是这种方法的延伸。在二战期间，美国农业部的一个社会科学专家组开展了一些与战争有关的社会调查，从那时起，区域概率抽样逐渐成为社会调查中一般总体抽样的主要方法。区域概率抽样还是进行家庭个人访谈调查选择样本的方法。康弗斯（Converse，1987）详尽记述了在美国发生的这一调查方法上的巨大改革。

从 1950 年起，抽样策略得到了极大的改进，最引人注目的进步就是通过电话随机拨号方式抽取家庭样本。无论怎样，良好抽样的实践原则经过长时间的发展建立起来了。

问题设计

调查过程中另一个重要的部分就是使用问题作为测量手段。最初的调查工作是新闻工作的延伸，不注意调查问题提出的方式。然而，其不足很快就表现出来：派出一名访谈员，带着一系列有问题目标却没有特别加以表述的问题去调查，结果获得的答案是五花八门的。因此，在20世纪早期，研究者们就开始编写标准化问题用来测量主观性现象。美国农业部的研究者在20世纪40年代广泛使用标准化问题来调查所需要的事实或客观资料，这方面的工作使他们获得了人们的赞誉。在佩恩(Payne，1951)的一本具有划时代意义的书里，提供了编写明确问题以使访谈员能够依照操作的实用指导原则。一般认为，李克特(Likert，1932)将心理物理学家开发的用以测量主观现象的复合量表技术(例如：Thurstone & Chave，1929)运用于实用社会调查研究的实践中。

在过去20年里，问题设计上的主要进步是改进了对问题进行评估的策略。与以前相比，研究者们现在都会测试问题以确定它们是否易懂以及答案是否有意义。通过分析被调查者的录音磁带来找出有疑问的问题，调查的先行测试已变得更为系统化。最终，问题表述的选择就具有了更多客观性而减少了调查人员的主观随意性。

访 谈

虽然并非所有的调查都包括访谈(因为一些调查只需调查对象自己填写问卷)，但是由访谈员提问并记录答案的方法仍是一种普遍使用的方法。当派访谈员进行访谈时，最重要的就是要避免他们影响被调查人回答问题，同时又要尽可能地增大被调查人回答问题的准确性。

要使访谈员始终保持一致性，首要方法是给他们标准化的问题，其次是要注意对访谈员的培训，让他们学会操作调查，避免在收回的答案中引入大量的偏误(Friedman，1942)。1954年，海曼(Hyman)、弗尔德曼(Feldman)和斯滕伯(Stember)出版了一套丛书，研究了调查员的报告方式与调查问题的表述对获得的问题答

案的不同影响。他们的工作引起对访谈员更精心的培训，培训内容是关于两方面的策略：一是探察何时会得到不完整的问题答案；二是如何正确处理访谈中的人际关系。1977年，坎内尔（Cannell）、奥克森伯格（Okesenberg）和康弗斯（Convers）进行了降低访谈员之间差异的尝试，他们把特别写好的自我介绍和一些鼓励语句交给访谈员，让他们在与被调查人交谈时使用，同时减少调查员与被调查人员之间无结构的交流。如今，访谈员的培训和监督对保证资料质量的重要性已得到了很好的证明（Billet & Loosveldt，1988；Fowler & Mangione，1990）。

资料收集方式

直到20世纪70年代，大多数学术性的和政府的调查都是采用面对面的访谈方式，由普通的访谈员进行。当拥有电话的人在美国越来越普遍时，电话访谈成为资料收集的主要方式。最新的资料收集方式是利用互联网。目前由于很多人缺乏上网的条件，它的使用十分有限，但是，上网条件发展得十分迅速，毫无疑问，互联网很快就会成为主要的资料收集方式。

整体调查设计

在20世纪50年代，良好的研究实践原则从各方面都有了相当的发展。可以理解的是，随着科学技术的发展，研究的过程和手段也发生了变化。在有些情况下，对于如何最有效地为特定的目的收集资料，我们缺乏研究。甚至一些调查实践已经被确认为十分完美了，但是其所使用的调查程序的质量仍具有易变性。

导致调查质量不稳定的原因众多。存在不精确的数据足以导致某些调查的质量不高；缺少资金、缺乏合格的工作人员、不具备一定的方法论知识，无疑都会使调查结果不尽如人意；还有关于严格概率抽样和标准化问题表述对调查质量价值的讨论（参见Converse，1987；Groves，1989；Suchman & Jordan，1990；Turner & Martin，1984）。调查研究设计时常受到批评的一点，就是研究者们不能保证上述三个主要工作领域中各个操作步骤具有较高质量。经常能看到研究者们非常注重良好调查设计方案的某些方

面,同时却忽视了另一些方面。本书的一个重要目标就是所谓的整体调查设计视角。

每一个调查都包括了许多能潜在增强或降低调查评估准确性(或精确性)的决策。总体来说,能让人获得更好的资料的决策包括更多的资金、时间和其他资源的投入,因此,一个调查设计就涉及一系列有效利用各种资源的决策。最理想的设计需要重视调查过程的所有重要方面。

关于抽样,关键性的问题有以下几个:

- 是否选择使用概率样本(probability sample)
- 样本框(sample frame,那些实际上有机会被抽样的人)。
- 样本大小。
- 样本设计(用来抽取人群或家庭样本的专用方法)。
- 回收率(样本中实际返回资料的百分比)。

关于问题设计,研究者必须判定关于问题的信度与效度的已有文献可被利用的程度、如何使用问题设计专家作为顾问以及在先行测试和问题评估中的投入。关于访谈员方面,研究者们可以对培训量、培训类型以及对他们的监管程度进行选择。考虑完这些方面之后,设计决策就开始决定资料的收集方式:是通过电话、邮件、个人访谈还是通过其他方式收集资料,并明确如何运用这些方式。电脑的运用也是决策要包括的。资料收集方式的决策将极大影响到所收集资料的质量。

这些片段合在一起就组成了整体调查设计。设计的所有组成部分通过保证调查资料的质量与调查设计的最容易出错的特征之间的关系联系起来。在过去,有时研究者们会关注到调查的某一两个特征,诸如样本规模和回收率,用来评估资料的可能的质量。然而,目前最好的调查研究是需要考虑所有上述的设计特征。如果在调查设计的任何一方面中有了重大的疏漏或缺陷,调查工作的其他方面的大量投入就会失去效力。例如,如果提问一些被调查人不能给予准确答案的问题,一个以最大限度减少抽样误差为目的的规模很大的样本就会失去意义。类似的甚至更常见的情况还有:如果样本设计不严密,或者回应率小到使样本失去了代表性的程序,或者调查员缺乏训练和监督,那么即使调查获得的回应者

的数量再大，也不能增加调查的可信度。

对于调查研究的设计者和使用者来说，整体调查设计方法意味着，当试图评估调查的质量和一套特定资料的可信度时，涉及的问题必须要包括以上提到的所有特征而不是其中的一部分。

本书的目的和目标

本书讨论了对任何调查研究进行设计的重要决策、研究者可进行的选择以及影响误差的数量和调查评估信度的不同选择的重要性，并提出了一套经过适当组织就可以构成良好调查实践的程序。本书着重探讨了研究者必须努力解决的一些现实和实践问题以及至关重要的理论上和方法论上的问题，以期使研究者能够克服很多资料收集上来源于错误的操作细节的缺陷，而不只是来源于通常意义上的那种不足。

显然，像这样相对简短的书不得不对内容进行某些选择。更全面地探讨诸如抽样、问卷设计以及访谈员的研究等专题的著作业已存在，要开展调查研究项目的读者应进一步阅读这些著作。此外，无论读这本书还是读任何其他的书，都不能代替在那些既有牢固方法论知识又有丰富的设计和施行调查经验的专家指导下的亲身实践和训练。然而，本书有一个它能胜任的重要角色，即它提供了全面的关于误差来源和调查资料收集中的方法论问题的总看法。

对于很多人来说，这些知识是适当和珍贵的。当然，社会科学家在他们的工作中使用别人收集的资料时也应该清楚地知道误差的来源。同样，阅读以调查为基础的统计数据的人也需要了解资料收集过程。本书明确了那些使用资料的人们会提出并需要得到回答的问题。另外，本书还为那些计划招标或委托调查的人提供了其所需的建议。总之，本书试图为那些将成为调查研究设计者和使用者的人提供了了解调查研究的新视角和新理解，同时也为那些实际上将要成为资料收集者的人提供一个完好的铺垫。



抽 样
Sampling

样本代表总体的程度是由样本框、样本规模以及抽样过程的特定设计决定的。如果使用的是概率抽样过程，样本估计值的精确度是可以计算出来的。本章描述了各种抽样过程以及它们对样本代表性和样本估计值的精确度的影响。本章更多叙述了两种最流行的人口抽样方法，即区域概率和随机电话拨号抽样。



有时，资料收集的目的不是为了获得有关总体的统计数据，而是要用更概括的方式描述一群人的特点。有时，新闻工作者、产品开发者、政治家等只需要大略了解人们的感情态度，而不太注意数字描述的准确性。如果研究者要进行小规模试验性研究以测量人们具有某种思想观点的广泛程度或者貌似一致实则不同的程度，那么那些便于找到的人（如朋友、同事）或自愿参与者（如杂志调查的应答者，电话公开的人们）就会显得很适用。并非所有的资料收集都需要严格的概率抽样调查，然而在多数情况下，进行调查的目的是为了展现有关总体的统计数据。本章关注的就是以产生统计数据为目标的抽样，这种抽样属于社会科学家可获得的各种形式的统计技术之列。虽然很多一般原则适用任何抽样问题，但是本章着重讨论的是人的抽样。

评估样本的方法不是通过结果和样本的特点，而是通过考查选择样本的过程。以下是样本选取的三个关键方面：

- 样本框是通过一定的抽样方法选定的一组有同样机会被抽取的人。从统计学上讲，样本只能代表包含在样本框中的总体。设计问题之一就是如何更好地使样本框与所要描述的总体保持一致。
- 概率抽样过程必须用来选择个体单元以组成样本。抽样过程必须使每个人都有被抽取的机会。如果研究者的处理或者回应者的特点影响到被抽取的几率，例如运用了便于找到的或主动自愿的回应者，那么就会失去评估样本代表性好坏程度的统计学的基础，通常使用的计算置信区间以评估样本的方法就不适用了。
- 样本设计的细节，诸如样本的大小以及选择单元的特定步骤等，都会影响到样本估计值的精确性，即影响到样本如何尽可能近似反映总体的特点。

这些抽样操作过程的细节，连同从被抽取的样本上能获得资料的比率，构成了评估一个调查样本质量所需的依据。本书第 3 章中将对回应率进行讨论，其中还将简要讨论定额抽样，它是对概率抽样的简单修改，这种抽样能产生非概率样本。本章将讨论抽样框、概率抽样过程，并介绍了几种行之有效的方法。对此有兴趣

的读者可以在基什(Kish,1965)、萨德曼(Sudman,1976)、卡滕(Kalton,1983)、格罗夫斯(Groves,1989)、亨利(Henry,1990)和洛尔(Lohr,1998)的著作中获得更多的相关知识。尽管准备进行调查的研究者经常会被建议去咨询抽样统计学家,但是,本章还是努力于让读者熟悉一些他们在评估调查抽样时会遇到的、应当予以关注的问题。

样本框

任何样本抽取过程都会让一部分人有机会被抽取到样本中,而排除另一部分有同样机会的人进入样本。样本框是由这两部分有机会被抽取组成样本的人构成。评估样本质量的第一步就是明确样本框。大多数抽样设计属于以下三个层次之一:

- 抽样是在相对完整的总体名单中进行的。
- 抽样是在一群到某处或做某事的人(例如:接受医生治疗的人或是参加某个会议的人)中进行的。在这种情况下,抽样进行之前是没有名单的,名单的产生和抽样过程是同时进行的。
- 抽样是分成两个或多个阶段进行的,其第一阶段是对其他有关方面的抽样,而不形成最终个体样本。通过一步或几步,实现基础的单元抽样,形成一份个体的或其他抽样单元的名单,再以此为基础实现最终的人员样本选择。一个最常见的这种抽样设计的例子是,在预先不知道哪些人居住在某些家庭单位的情况下,先选择一些家庭单位,然后再选择住在这些家庭单位里的人员样本。本章将在后面详细阐述这种分阶段的抽样过程。

研究者应该评估样本框的以下三个特点:

1)全面性:一个样本只能代表一个样本框也即是有机会被抽样的总体。大多数抽样方法总会从所研究的总体中排除一些人,例如,以家庭为基础的样本就排除了那些住在诸如集体宿舍、监狱、疗养院等集体住处的人,也排除了那些无家可归的人。便于获

得的常见人员名单，诸如有驾照的司机、登记的选民、房屋所有者等将更具有这种排除性。虽然它们包括了某些总体的大部分，但却排除了重要的有独特特征的个体。一个典型的例子：公开的电话号码簿就排除了那些没有电话的人、要求不公开其电话号码的人以及在最新电话簿出版后才安装电话的人。在一些中心城市，几乎有50%的家庭会这样被排除掉。由此，从这些城市的电话簿里得到的样本只能代表大约一半的总体，并且可以断定，这一半会在很多方面不同于另一半。

电子邮件地址抽样提供了另一个典型的例子。对于有的总体，如在公司或学校的人，使用电子邮件就很方便。而对一般家庭总体使用电子邮件地址的抽样办法，就会排除很多人，产生的样本就会在许多重要方面与总体有很大的不同。

评估任何抽样设计的关键，是确定有被抽取机会的个体占总体的百分比及其与没有被抽取机会部分的差别程度。研究者会经常面临两难的选择：一个是简单便宜但会排除了一些人的方法；另一个是花费多但抽取的样本能更全面代表总体的方法。如果研究者想从已有的名单里抽样，特别重要的是要仔细评估这份名单以确定它是如何编制的、它的增删是何时及如何进行的，并要确定可能被排除在名单外的人员的数量和特点。

2)选择的概率：有可能计算出每个人被抽取的几率吗？在一年内就医记录的抽样中，就医次数多的人比只有一次就医经历的人有更高的被抽样几率。抽样设计没有必要给予抽样框中的每个人(如在名单中出现且仅出现一次的那种情况)相同的被抽取的概率。但重要的是研究者能够算出每个个体被抽取的概率，这些应该在通过核查抽样人员名单以选取样本时进行，也有可能在资料收集时进行。

在上面提到的通过就医次数来抽取病人样本的例子中，如果研究者询问被抽取的病人一年就医的次数，或者研究者能获得被抽取的病人的就医记录，都有可能在分析被抽样概率的基础上校正调查资料。但是，如果研究者不可能知道每个人被抽样的概率，也就不可能准确评估出样本统计数据与其总体之间的关系。

3)有效性：在有些情况下，抽样框(sampling frame)包括了研

究者不想研究的抽样单元，如果在资料收集时能识别出符合要求的个体，那么过于全面的样本是无可厚非的。因此，抽取家庭中老年人样本的最好的方法是将所有家庭都归入样本框中，找出那些有老年人的家庭后，再排除没有老年人的家庭。随机拨号抽样是以抽取电话号码（有些并未使用）作为抽取拥有电话的家庭样本的方法，这些设计的惟一问题是它们是否合算。

因为从样本中进行概括的能力是由样本框所决定的，当研究者公布其结果时，他必须告诉读者哪些人被抽样哪些人被排除，以及被排除的人的特异程度。

抽取一阶样本(one-stage sample)

研究者决定了样本框或获得样本的途径之后，就应该决定怎样抽取包括在其中的个体单位。下面几节内容将讨论抽样者抽取样本的各种典型方法。

简单随机抽样

从某种意义上来说，简单随机抽样是典型的从总体中进行的抽样。所有计算样本统计数据的最基本方法都是假定抽取了一个简单随机样本。简单随机抽样类似于从帽子中摸出一个样品：一次选择一个总体的成员，每个成员是独立且不能互相代替的；一旦某个单位被抽取，他就没有再被抽取的机会。

从操作上来讲，进行简单随机抽样需要一个编好号的总体的名单。只要确定每个人在总体中存在并只出现了一次就行了。假如一个名单上有 8 500 人，简单随机样本只要 100 个人，那么过程就很简单了：对名单上的人从 1 到 8 500 进行编号，然后使用电脑、一张随机数表或者其他产生随机数字的装置，就可以在同样的范围内产生 100 个不同的数字，与数字相对应的 100 人就组成了人数为 8 500 的总体的一个简单随机样本。如果名单是电子档案形式的，就先把名单随机排序，然后选择新名单上的前 100 人，这样也会产生同样的结果。

系统抽样

如果名单很长，所有的抽样单元都未先编好号，或者不能方便地用电脑操作进行编号，那么上述简单随机抽样的方法用起来就会十分吃力。在这种情况下，可以使用另一种抽样方法，即系统抽样。系统抽样与简单随机抽样具有相同的精确性，并且操作方便。另外，系统抽样便于使分层的优越性体现出来（以下部分将予以讨论）。

如果要从一个名单中选出系统样本，研究者首先要决定名单上项目的数量以及要抽样的项目的数量，然后用后者除以前者。这样，假如名单上有 8 500 人，样本只需 100 人，名单人数的 100/8 500会被包括进样本（即从每 85 个人中选出 1 个）。先在样本间距里选择一个随机数，在本例中就是从 1 到 85 中的任意一个数字，随机定的起点保证了它是一个随机抽样过程；从随机最初选出的那个人起，研究人员就可以按名单每数 85 个人就挑 1 个出来。

大多数统计学著作都提请注意，如果名单是按某些特征排序或者有一种循环格局的情况时，即使是随机的起点，系统样本也会受到明显影响。有一个极端的例子，在一个夫妇俱乐部的成员名单里，如果丈夫的名字总是排在妻子前面，那么只要是任何偶数的间距，总会产生一个只有一种性别的系统样本。在检测一个可能的样本框时，注意到由一个随机起点产生的样本是否会与由其他随机起点产生的样本对调查结果有系统的不同影响，是十分重要的。在实践中，大多数名单或样本框不会产生不适合于系统抽样的问题。果真问题出现时，只要调整名单或抽样间距，通常情况下是可以达到至少可与简单随机样本相媲美的系统抽样效果的。

分层抽样

当抽取简单随机样本时，所抽取的个人是独立的，不受先前已经抽取的其他个人的影响。这一过程有时会导致样本的一些特点与其总体不同。一般说来，在资料收集以前，对总体中个体的特点知之甚少。然而，有一些总体的特点经常在抽样时就能得到确认。在这种情况下，有可能重新计划抽样程序以减少常规抽样的变异，

以便形成一个比简单随机样本更接近于总体特征的样本。达到这个目的的过程被称作分层。

例如，某研究者有一个大学生名单，其中姓名是按字母排序的，名单里不同班级的学生名字混杂在一起。如果名单标明了每个学生所属的班级，就有可能重新将名单排序：把大一学生排在最前面，接着是大二、大三，最后是大四的学生，并且所有的班级都组合在一起。如果抽样设计要求在名单中每 10 个学生中选出 1 个，重新排序后的名单就能保证有 1/10 的大一学生被抽样，有 1/10 的大二学生被抽样，以此类推。如若不然，而是用原来的字母排序名单形成简单随机样本或系统样本，那么大一的样本比例就由标准抽样变异决定，它会比其在总体中的实际比例更高或更低，而提前分层就确保了抽样比例与实际比例相同。

现在来考虑一下评估学生平均年龄的工作。由学生组成的班级几乎可以肯定是与年龄有一定关系的。尽管由于抽样过程的原因，在样本评估中会存在一些可变的因素，但是在抽样框中注意班级的代表性，可以抑制样本平均年龄偶尔出现与总体有差异的程度。

几乎所有以地理区域为总体的样本都可以依据某些区域变量进行分层，这样做可以使样本的分布方式符合总体的特征。国家样本是典型地通过国家的不同区域或按城市、市郊和乡村分层进行抽样的。虽然分层只增加了与分层变量有关的变量评估值的精确度，但是由于某些层次的分层相对来说简单易行，并且它从不会降低样本评估值的精确性（只要所有层次中所用的是同样的概率抽样），因此分层抽样通常是样本设计的一个亮点。

抽取的不同概率

有时分层被用作改变总体中各种子总体抽样率的第一步。当各个层次的抽样概率基本一致时，一个占总体的 10%的子总体，也会构成样本的 10%。如果研究者想从占总体 10%的子总体中选出至少 100 人的样本，那么简单随机抽样法需要一个总数为 1 000 人的样本；如果研究者打算将子总体的样本规模扩大到 150 人，那么就需要追加 500 人到总的样本中，使总的样本规模达到 1 500 人，这样才能使总样本的 10%等于 150。

显然，有时以这种方式增加样本既费时也费力，而且也不怎么有效。在上面的例子中，即使研究者对其他子总体的样本规模满意，但为了增加 50 个有用的调查对象设计一个子总体样本时，也不得不同时增加 450 个不需要的调查对象以备抽取。因此，在有些情况下，合理的设计方法是让某些子总体有比其他的子总体拥有更高的抽样率。

例如，假设研究者想在一所男生仅占 20%的大学里，却至少要用 200 个男生来对比男女学生的情况。这样，500 人的样本中应该有 100 个男生。但是如果能事前区分出男生，就可以将男生的抽样概率提高为女生抽样概率的 2 倍。这样不用再加 500 人以从中再选出 100 个男生，而可在原来的那 500 人样本基础上另增加 100 个男生，组成 200 人的男生样本。由此，当进行男女生比较时，就会得到 200 男生和 400 女生作为调查对象，其精确度不会下降。为了组合这些样本，研究者应给予男生的权值为女生的一半，因为男生的抽样概率是女生的 2 倍。

表 2.1 不同概率抽样示例

	女生	男生
总体的人数	4 000	1 000
在总体中所占比例	80	20
抽样比例	1/10	1/5
样本中的人数	400	200
样本里没有加权处理的比例	66.7	33.3
加权(以调整概率选择)	1	1/2
样本里加权处理后的人数	400	100
样本里加权处理后的比例	80	20

在表 2.1 里，可以看到被调查者中有 33.3%的是男生(600 人中有 200 个男生)。但是采取了加权的办法后男生组成了加权后样本的 20%，这就与他们在总体中的所占比例相等。

即使对所关注的子总体中的个体不能在抽样前了解清楚，有

时也能应用这里提到的基本方法。例如，一般不可能在接触前就获得确认了居住者种族的家庭名单。然而，亚裔家庭通常比其他种族的人居住更为集中。这种情况下，研究者就可以用高于平均数的比率抽取亚裔占优势地区的家庭，以增加亚裔调查对象的数量。再次强调，任何子总体与总体中的其他子总体的抽样率不同时，为了使组合的或总的样本产生关于总体的准确的统计值，就需要进行适当的加权处理。

第三种方法是在与有可能成为调查对象的人接触而收集到的资料基础上调整抽样率。仍以上述大学生调查为例，假设事先不能查明学生的性别，研究者可以选择一个有 1 000 名学生的初始样本，让访谈者把每个学生的性别查明，然后对所有被选的男生(200)和一半的被选女生(400)进行访谈。其结果也会与刚才谈到的表格里的结果一样。

最后，应该指出在不同层面使用不同抽样率的另一个技术上的原因：如果一个组群在所测量内容上与另一个组群有很大的不同，那么，相应地对这个组群进行过度抽样(over sample)，将有利于提高最终整体评估值的精确性。格罗夫斯(Groves，1989)很好地阐述这一基本原理，并提供了评价这些设计的效度的方法。

多阶抽样

当没有合适的总体成员名单，并且没有办法直接把握总体时，多阶抽样就成为一种很有用的抽样方法。

没有可直接抽样的来源，就需要一种将总体中的成员与某种能够进行抽样的分组联系起来的策略。第一阶段先对这些分组进行抽样，然后再由所选定的这些组的成员形成名单，以便在后续阶段能从这些名单中再进行抽样。在抽样术语学上，抽样设计的第一阶段里分出的那些组通常被称为“整群”(clusters)。下面通过介绍在名单不可得的三种最普通的情况下对多阶抽样的使用，说明多阶抽样的一般策略。

从学校里抽取学生样本

要从某城市的公立学校中所有注册在读的学生里抽出一个样本，却找不到这些学生的一个完整名单，是常有的事情。然而有一种抽样框可以让研究者把握所需总体的所有学生的情况：即关于那座城市里所有公立学校的名单。因为研究总体中的每个学生都只与一个学校有联系，所以关于学生满意的完整样本就可以通过使用二阶策略抽样获得：首先选择学校（即整群），然后从那些学校里选择学生。

假设有以下资料：

城市里有 40 个学校共有 20 000 名学生
目标样本＝2 000＝1/10 的学生

那么将有如表 2.2 所示的四种抽样的设计或方法，其中每一种都会产生 2 000 人的概率样本。

表 2.2　多阶抽样示例

	第一阶段的选择概率（学校）	×	第二阶段的选择概率（在所选学校里的学生）	＝	总的选择概率
1）选择所有学校列出所有学生，然后在每个学校选择 1/10 的学生	1/1	×	1/10	＝	1/10
2）选择 1/2 的学校，然后在其中选择 1/5 的学生	1/2	×	1/5	＝	1/10
3）选择 1/5 的学校，然后在其中选择 1/2 的学生	1/5	×	1/2	＝	1/10
4）选择 1/10 的学校，然后收集其中所有学生的资料	1/10	×	1/1	＝	1/10

以上四种方法都将产生 2 000 人的样本，都给予每一个学生同样的(1/10)抽样概率。四种方法的不同在于：从第一种到最后一种，越来越节约经费，获得名单所需的学校越来越少，需要走访的学校越来越少；同时，样本的精确度会因抽取的学校更少和每所学校里抽样的学生更多而降低。有关诸如此类的多阶抽样设计对样本评估值精确度的影响，将在本章后面进行更详细的讨论。

区域概率抽样

区域概率抽样因其广泛适用性而成为多阶抽样策略中最常使用的方法之一。它可以用来抽样以地理区域性划分的任何总体，诸如住在某个地区、某个城市、某个州或某个国家的人们。基本方法就是将所有对象区域划分为全面彻底、互相排斥、界限明确的子区域，这些子区域就是整群，由子区域组成的样本也就此产生；再由选定的子区域里的家庭单元组成一个名单；然后，抽取写在名单上的家庭单元组成样本；最后一步，是用抽取的家庭单元的所有成员组成样本，或者进一步用这些成员的名单再次进行抽样。

这种方法适用于丛林、沙漠、人烟稀少的乡下或大城市的中心区。抽取这种样本的详细步骤可能十分复杂，但是，通过描述如何以城市街区为子区域单位抽取城市总体样本能够阐明其基本原则。假设有以下资料：

一个城市有 400 个街区
共有 20 000 个家庭单元住在这些街区里
目标样本＝2 000 个家庭单元＝1/10 的所有家庭单元

考虑到这些信息，可选择与从学校抽取学生样本相似的方法去抽取家庭样本。抽样的第一个阶段，选定街区(即整群)，如表 2.3 所示。在第二个阶段，列出被选定街区里的家庭单位名单，从名单里抽取样本。接下来介绍两种抽取家庭单元的方法。

表 2.3 区域概率抽样示例(一)

	第一阶段的抽样概率(街区)	×	第二阶段的抽样概率(选定街区的家庭单元)	=	总的抽样概率
1)选定 80 个街区(1/5),再选出这些街区里 1/2 的单元	1/5	×	1/2	=	1/10
2)选定 40 个街区(1/10),再选出这些街区里所有的单元	1/10	×	1/1	=	1/10

与学校抽样相类似,由于第一种方法比第二种所选定的街区更多,因此,成本更高,能获得比规定规模更精确的估计值。

没有任何一个设计在选择抽样方案时考虑到了第一阶段的组群规模(例如,街区或学校的规模)。规模大的和规模小的学校或街区有同样的抽样率,因此,在最后一个阶段,每个被抽样的组群都有一个固定的比例,那么与小的学校或街区相比,大的学校或街区的访谈的数量将更多。在第二阶段,样本的规模(整群规模)就会有相当大的不同。

如果有关于第一阶段抽取的组群规模的资料,那么对这些资料予以运用往往是有益的。如果在抽样的最后一步中选定的单元数量在所有的整群中都相同,那么样本的设计将产生更精确的估计值。这种设计方案的优点还在于能更容易计算出抽样误差,并且更容易预见样本的总体规模。为了产生同等规模的整群,在第一阶段进行单元抽样时,要与它们的规模成比例。

接下来的例子展示了根据街区规模来按比例进行抽样的方法。它是以区域概率法作为对家庭单元(公寓或个人住房)进行抽样的第一步。这种方法也适用于学校抽样,只要把学校当作街区就可以了。

1)确定在抽样的最后阶段要抽取的家庭单元的数量,即平均

的整群规模，例如选 10 个。

2）对第一阶段抽取的每个街区所包含的家庭单元数量作出估计。

3）给街区排序，以便把地理位置邻近或相似的街区排在一起，这是为提高样本的质量而进行的划分。

4）产生所有街区家庭单元的评估累计数，最终将形成表 2.4。

表 2.4　区域概率抽样示例（二）

街区号	家庭单元估计数	家庭单元的累计数	得分（随机起点＝70，间距＝100 个家庭单元）
1	43	43	—
2	87	130	70
3	99	229	170
4	27	256	—
5	15	271	270

5）决定整群的间距。如果要从 10 个家庭单元中选出 1 个，并且在每个选定的街区中选择一个由 10 个家庭单元组成的整群，就需把整群的间距定为 100 个家庭单元。换种方法，我们不从每 10 个家庭里选出 1 个家庭，而从每 100 个家庭里选出 10 个家庭，选择率是一样的，但是使用的方式是“整群的”。

6）在从 1 到 100（本例的间距）随意选择一个数字作为起点，进而按照累计数系统地进行，标明在第一阶段抽样的最初的单元数（街区）。在本例中，随机抽样的起点（70）就跳过了第 1 街区（有 43/100 的被抽中的概率），第 70 个家庭单元在第 2 街区，第 170 个家庭单元在第 3 街区，第 270 个家庭单元在第 5 街区。

由被选定的街区（第 2，3，5 街区）的家庭单元组成的名单，通常是派人对这些街区进行走访形成的。接着就是从这些名单中选择出家庭单元。如果能确保对这个街区规模的估计值是准确的，那么研究者就可使用简单的随机抽样或系统抽样在每个选定街区

中选出 10 个家庭单元。一般认为系统抽样是最好的，因为由此选定的单元会分布在整个街区。

对第一阶段抽样单元（例如街区）规模的评估通常有误差，可以通过计算从街区中所选择的家庭单元的抽样率来修正这些误差，例如：

街区家庭抽样率＝整群的平均规模/街区家庭单元的估计数
＝10/87＝1/8.7（在第 2 街区里）

本例中，在第 2 街区，我们可以在每 8.7 个家庭单元里抽取 1 家；在第 3 街区，每 9.9 个家庭单元里取 1 家；在第 5 个街区，每1.5 个单元取 1 家。如果某街区规模比预想的大（例如，进行了新的建设），所抽取的样本就不止 10 家；如果它比预期的小（例如拆迁），所抽取的样本就少于 10 家；如果与我们预想的一样（例如，第 2 街区中有 87 个家庭单元），我们就选 10 个（87÷8.7＝10）家庭单位。通过这种方式，就可以在抽样的过程中自动修正最初对街区规模评估的误差，同时使得所有街区中的家庭单元都有同样的被选择的机会。不管是估计的还是实际的街区规模，任何家庭被选择的概率都是 1/10。

区域概率抽样法可以用来对任何以地域划分的总体进行抽样。虽然区域越大，步骤就会越复杂，但是方法都是一样的。要记住以下的关键步骤：

- 所有区域都必须给予一定被抽取的机会。将预计没有任何符合条件的抽样单元的区域与相邻的区域结合起来以保证获得抽样机会。在这个过程中可能会因为有了新的建设或评估本身的偏误而产生误差。
- 街区（或其他类的地区）的抽样概率与从样本街区里抽取家庭单元的概率之积，在所有街区应该保持相同。

最后，即使是细心的现场制表人也会漏掉一些家庭单元。因此，在资料收集时注意核查、防止遗漏是必要的。

随机电话拨号

随机电话拨号(RDD)是另一种抽取家庭单元以便抽取这些家庭成员样本的方法。假设有 6 个电话局覆盖了 20 000 个家庭单元,研究者可以从拥有电话的 10%的家庭单元中抽取概率样本:

- 在 6 个电话局里,大概有 60 000 个电话号码(每个电话局有 10 000 个)。选出 6 000 个(即 10%),也就是从每个电话局中任意抽取 1 000 个 4 位数的电话号码。
- 拨打 6 000 个号码。并非所有的号码都是家庭电话号码,而且还有许多无用的号码、连接不通、暂时无法使用或是商业电话。由于该区域内有 10%的电话都被打通,因此,在这个地区大约 10%的拥有电话的家庭能够通过拨打电话进行联系。

这就是最基本的随机电话拨号抽样方法。此法最明显的不足就是有大量的无用电话。在全国范围内,最多有 25%的是家庭住宅号码,在城镇地区大概有 30%,在农村大概只有 10%。瓦克斯伯格(Waksberg,1978)建立了一种依据电话号码分组的方法。每组包含 100 个电话号码,每个电话号码由 3 个区位号码,3 个电话局的号码和 2 个附加码(区位号码-123-45XX)组成。首先,在一个样本中任意拨打号码以便对其进行初步筛选;接着,便在家庭住宅号码组中再任意拨打,这样家庭单元被抽中的概率就可以被提高到 50%以上。在这个设计中,100 个电话号码构成的组就是“整群”。

近些年来,大多数调查机构开始使用名单辅助法进行随机电话拨号。随着电脑技术的发展,公司可以编辑电子电话簿,这些电子电话簿每 3 个月更新一次。一旦所有的电话簿都是以电子文件的方式存在,就可以搜寻到至少有一个公开(出现在电话簿上的)住宅电话号码的所有整群(区域码-123-45XX)。这些公司就可将这些在组群中至少有一个公开住宅电话号码的所有电话号码组成一个样本框,然后利用这个样本框进行抽样。这种方法有两个优点:第一是瓦克斯伯格方法里所要求的电话号码的最初的删筛工作就不再需要了,因为编好的样本框已经完成了这项工作;第二是使用这个样本框选好的样本不再需要分组了,通过使用包含了所有的住宅电话号码整群的样本框,就可对电话号码进行简单随机

抽样了。这种新方法比以前的方法更加经济有效。其局限在于，在整群中，没有被列出的住宅号码就没有被抽中的机会。布里克，瓦克斯伯格，库尔普和斯塔尔(Brick，Wakesberg，Kulp，Starer，1995)估计在美国平均有4%的有电话的家庭不在名单内。列普克夫斯基(Lepkowski，1988)对以家庭抽样为目的的各种电话号码抽样方法进行了详细的总结。

另外，在使用随机电话拨号抽样方法时，还应该注意一些问题。首先，它的使用价值是建立在大多数家庭有电话的事实基础之上的。在全国范围内，大概只有5%的家庭没有电话，但是在一些地区，特别是中心城市或乡村地区，遗漏率可能更高。桑贝里和马西(Thornberry & Massey，1988)对那些有和无电话的家庭的不同做了进一步的分析。

随着手机使用率的上升，家庭座机使用率下降，RDD抽样将面临一个问题：普通的RDD抽样只关注家庭座机而避免对使用手机的家庭进行抽样。对两种方式都进行抽样是可行的，但是假如将手机抽样加入抽样框，抽样、资料收集和调查后的加权就会变得十分复杂。

对RDD抽样提起另一个挑战是，如何找到在与地区边界不相符的电话局内进行小地区抽样的方法。除非能确保电话号码与研究区域相一致，否则访谈员就需要回应者告诉他们是否住在调查区。当回应者住在一个难以界定的地区时(例如，邻近地区)就很难保证抽样的可靠性。

像其他抽样方法一样，对所有的调查而言，随机电话抽样并不是最佳选择。对其赞成和反对的观点将在第4章进行详细讨论。无论怎样，在过去的30年里，作为抽样方法之一的随机电话抽样方法的引入，为调查研究水平的提高作出了重大贡献。

选择回应者

区域概率样本和随机电话拨号样本都是对家庭单位进行抽样，因而，下一个问题就是由家庭里的哪些人来接受访谈。

最佳的选择取决于所收集信息的种类。在某些研究中，要对家庭情况及其所有成员的信息加以收集。如果要求回答的问题很

简单，那么就可以选择这个家庭里的任何成年人进行回答；如果要求所做的回答是专门的，访谈员就应该对家庭中知识最渊博的人进行访谈。例如，在国家健康访谈调查里，家庭里"最了解家庭健康状况"的人将被作为涉及家庭中所有成员的问题的回应者。

但是，很多人都只能说出关于他（她）自己的事来。大多数研究人员普遍认为没有任何人可以说出他人的感受、观点和知识。还有一些通常只有自己才能准确说出的行为和经历（例如：吃喝的东西、所买的物品、看见或被告知的事物）。

当研究包括的变量只有通过自我讲述方式才能获得时，抽样过程除了选定家庭之外，还必须从这些家庭中抽取特定的个人。一种方法就是与家庭中的每个符合条件的人进行访谈（此时，不进行任何抽样）。除了前一个回应者的答案会影响后一个人的回答之外，由于整个家庭具有同质性，最常用的方法就是每个家庭只选择一个回应者。很显然，选择那些恰巧接电话或开门的人作为回应者是一个非概率抽样和存在潜在偏误的选择样本的方式。访谈员的处置权、回应者的处置权以及有效性（与工作地位、生活方式和年龄相关）都会影响到谁将成为回应者。概率抽样的重要原则就是通过随机抽样来选定具体的人。在家庭内运用概率选择回应者的过程包括以下三个步骤：

- 明确家庭里有多少人适合做回应者（例如，有几个 18 岁或 18 岁以上的人）。
- 对所有家庭以同一种方式进行编号（例如，按降序排列年龄）。
- 要有一个客观的选择回应者的过程。

基什（Kish，1949）编制了一个通过使用随机调查表选择回应者的详细步骤，此表至今还在使用。当有电脑辅助访谈时，让电脑来选择合适的家庭成员是很容易的。这个过程重要的特点是没有人为的处置权的干扰，在被选定家庭里，所有符合条件的人都有被选中的可能。格罗夫斯和利伯格（Groves & Lyberg，1988）对几种简化回应者选择过程的策略做了评论。

当一个家庭只访谈一个人时，则需要引入选择的差异率概念。如果一个家庭只有一个成年人，很显然当选定这个家庭时这个成

年人将会成为回应者;如果一个家庭有 3 个成年人,每个成年人就只有 1/3 的机会被抽样。无论何时,为了避免被抽样的人的资料在样本统计数据中过多出现,就必须对一个以不同于其他组的比率进行抽样的那个组进行加权处理。在本章的前半部分有个例子,当男生的抽样率是女生的两倍时,他们的回答就得乘以权值 0.5,以便他们的被加权的样本的比例与总体中的比例相同。当从具有不同数量的符合要求的人的家庭中选择一个回应者时,可以使用同样的方法。

调整对每个家庭选择一个回应者的影响的最简单方法是根据那个家庭中符合条件的人的总数给每个答案加权。因此,如果有 3 个符合条件的成年人,那么权值为 3;如果有 2 个成年人,权值就为 2;如果只有 1 个符合条件的成年人,权值就为 1。如果加权方案是合适的,那么对于所有的回应者来说,抽样概率乘以权值的结果也是一样的。

还有一些变量与家庭中成年人的数量紧密联系。例如,结婚的人比单身的人更有可能会住在至少有两个成年人的家庭里。已婚者与单身者在很多方面都有所不同,因此假如数据不被进行加权处理,那么任何与婚姻状况相关的评估都将是不准确的。

也有一些变量与家庭里的成人人数没有关系,在这种情况下,加权也不会对描述性的结果产生影响。当需要产生描述性统计数据时(例如均值或分布),一般通过加权来调整家庭和/或成人抽样的差异率。此外,为描述性目的而进行的加权对很多标准分析程序包来说是十分容易的。但是在加权资料基础上使用统计数据测试的过程会十分复杂。在统计数据测试中用到的估计值除了其他的样本设计特点外,还与实际回应量相联系。如果对一个项目的观察比它实际上应得到的更多或更少,使用权值可能会使运算出错。有时当调查的设计很复杂时想要准确地进行统计数据测试是很困难的。而且对一些相关分析,没有必要进行加权(Groves,1989)。这是需要抽样统计学家对调查资料进行合理设计和分析的重要性的又一个例证。

从样本和抽样误差中得到估计值

以上提到的抽样策略被选择的原因是它们的使用极其普遍，并且显示了抽样设计选择的主要取向。一个概率抽样方案最终会在没有研究者或回应者的处置权的情况下选出一组家庭或个人。研究者所能使用的基本工具就是简单随机抽样和系统抽样，它们都可以通过分层、不等的抽样率和整群对其加以完善。抽样策略的选择部分地是由其可行性和成本所决定的，同时也包括了对样本估计值精确性的考虑。使用概率抽样方法的主要原因是可以使用各种统计工具去评估样本估计值的精确性。在这一部分，将对样本估计值的计算方法及其受样本设计特点影响的方式进行探讨。

研究者常常对样本本身的特点不感兴趣，收集样本的资料是为了得到反映总体的结论。本章对统计和设计问题的讨论，就是以探讨样本特征能精确描述总体状况的程度为出发点的。

有许多形象化的描绘被用来解释概率理论。可能其中最容易理解的就是掷硬币的方法。如果一个质量均匀的硬币被掷出了10次，不一定恰好是5个正面和5个反面，有时是6个正面，偶尔是7个正面，甚至还会出现10个正面。本质上讲，你可以把10次掷硬币的活动认为是许多可能的样本。如果完成一系列的10次掷硬币活动，在掷了10次后记下硬币正面出现的次数，再掷10次再记录下硬币正面出现的次数，一直这样做下去，就会产生一个分布图。如果是质量均匀的硬币，10次掷硬币得到正面的次数，更多的是5次。围绕5次正面向上就会出现一个分布图，10次向上或一次都不向上的极端状况出现的频率最低。

虽然调查中有些误差来源是由于偏误并产生了系统性的数字失真，但是抽样误差一般是由抽样的随机性引起的（因此不是系统性的偏误）。当用概率抽样法选取样本时，可能能够计算出由于抽样而引起样本估计值偶然变化的幅度。

如果对抽取的样本的数量没有限制，描述性统计的样本估计

值(例如,均值)就会正常分布在真实的总体值的周围。样本规模越大、测量内容的变动性越小,样本估计值就越能集中分布在真实的总体值周围,并且以样本为基础的估计值就会越精确。围绕真实总体值的变化(源于样本和总体之间的偶然差异)被称为“抽样误差”。针对随机抽样的易变性,确定样本评估的信度是评估调查数据质量的重要组成部分。

选择样本的设计(尤其是它是否含有分层、整群或不等概率抽样)会影响对一定规模的样本抽样误差的估计值。但是,通常描述抽样误差的方法,是看在简单随机抽样中它们可能以什么形式出现,然后计算出简单随机抽样设计中离差的影响。因此,首先要介绍简单随机样本抽样误差的计算。

简单随机样本中的抽样误差

虽然本书不是关于抽样统计的教科书,然而,对某个样本设计误差的评估却是调查设计过程的基本组成部分,并且研究者通常向读者提供关于把误差归因于抽样的一些准则,这些准则是专业读者和调查研究资料的使用者都应该知晓的。因此,对计算抽样误差方法的认识是了解整个调查过程的必要组成部分。

尽管同样的逻辑适用于一个样本的所有统计数据,但是最常用的样本调查估计值是均值或平均数。经常用来描述抽样误差的统计数据被称为标准误差(一个均值),它是按一定规模抽取的巨量样本的均值估计值分布的标准离差。标准误差值一旦确定,人们就可以说一定规模和设计的样本均值 67%是在真实的总体均值±1 的标准误差范围内,有 95%会在真实总体值±2 标准误差周围内。后面一个数字(±2 标准误差)通常被称为样本估计值的“置信区间”。

对一个均值的标准误差的评估是通过计算方差和被评估的样本规模得出的:

$$\mathrm{SE} = \sqrt{\frac{\mathrm{Var}}{n}}$$

SE——均值的标准误差

Var——样本离差的平方之和，即方差

n——样本大小

从样本调查中计算出来的最常见的均值可能是一个比值，即有某种确定特点或给出了某种回应的样本的百分数，它在表明两个数值分布均值的比例上很有用途。

均值就是平均数，它的计算方法是将总值除以事例总数：$\sum X/n$。现在假设只有 2 个值 0(否)和 1(是)，某样本中有 50 个成员，当问到是否结婚时，有 20 个人说是，其他的说否。如果有 20 个“是”和 30 个“否”的回答，就可这样计算均值：

$$\sum X=(20\times 1)+(30\times 0)=20;\quad \sum X/n=20/50=0.40$$

一个比例陈述，例如有 40%的回应者结婚了，恰恰是关于 1 或 0(即回答“是”或“否”，1＝是，0＝否，译者注)的分布均值，均值为 0.40。可以根据通过 $p\times(1-p)$计算出的比例的方差，然后很容易地计算出比例的标准误差，p 即是具有某种特征的部分(例如，例子中有 40%的人已婚)，$(1-p)$即是样本中没有某种特征的部分(例如，60%的人未婚)。

我们可以看到均值标准误差的公式：

$$\mathrm{SE}=\sqrt{\frac{\mathrm{Var}}{n}}$$

因为 $p(1-p)$是一个比例的方差，所以比例的标准误差计算公式变为：

$$\mathrm{SE}=\sqrt{\frac{p(1-p)}{n}}$$

在前面例子中，50 人中有 40%的人已婚，其标准误差的计算方法如下：

$$SE=\sqrt{\frac{p(1-p)}{n}}=\sqrt{\frac{0.40\times 0.60}{50}}=\sqrt{\frac{0.24}{50}}=0.07$$

因此，我们可以估计真实总体数(整个总体中已婚比例)在0.33到0.47(0.40±0.07)之间的概率是0.67(即样本均值±1个标准误差)；有95%的把握说总体的真实值在样本均值的两个标准误差之间，即在0.26和0.54(0.40±0.14)之间。

表2.5是关于不同规模和不同比例的样本抽样误差的概括性表格，表中所提供的样本是通过简单随机抽样产生的，表里出现的所有数字代表了一个比例的两个标准误差。考虑到对做了某种回应的样本比例的估计，表格为不同样本规模提供了95%的置信区间。在例子中，50个人中得出了有40%的人已婚的样本估计值，正如我们所计算的，表格提供的置信区间在0.14左右。如果100个人里得出了20%的人已婚的估计值，表格显示我们有95%的把握说真实值是在20%±8%的区间内(即12%到28%)。

表2.5　抽样变异导致的不同置信区间*

	有某些特点的样本的比例				
样本规模	5/95	10/90	20/80	30/70	50/50
35	7	10	14	15	17
50	6	8	11	13	14
75	5	7	9	11	12
100	4	6	8	9	10
200	4	6	8	9	10
300	3	4	6	6	7
500	2	3	4	4	4
1 000	1	2	3	3	3
1 500	1	2	2	2	2

注：在给定有特点的样本的比例以及产生这个比例的样本的规模的情况下，概率是95%的实际总体值取决于在表中值域的增减数。

* 本表描述了抽样的易变性。源于无回应或报告错误的误差在本表中没有反应。此外，本表假定所采用的是简单随机抽样样本。由于样本设计或其他方面的影响，估计值或许比本表所反应的值更具有易变性。

表中提到的几点都要加以注意。首先,可以看出样本越大产生的抽样误差越小。第二,也可以发现,当我们对样本的数量进行补充时,能减少样本误差,规模小的样本比规模大的样本误差的降幅要大。例如,将 50 个成员加入规模为 50 的样本中,就会使误差明显降低。但是将 50 个成员加入到规模为 500 的样本中,它对增加整个样本估计值的精确度所起的作用几乎看不出来。

第三,可以看到,抽样误差的绝对值在比例为 0.5 左右时最大,当具有某种特点的样本比例达到 0 或 100%的时,抽样误差的绝对值会下降。可以看到标准误差与方差有直接联系。当比例越远离 0.5 时,方差 $p(1-p)$就会越小。

第四,表 2.5 和它所基于产生的公式适用于简单随机抽样所产生的样本。一般总体的大多数样本不是简单随机样本,样本设计对计算抽样误差的影响程度,会因同一调查里的不同设计和不同变量而有所不同。通常,表 2.5 会对普通总体样本的样本误差做出保守的估计。

最后,要强调的是,表 2.5 中所描述的是源于抽样的潜在误差的易变性反应,而与总体中每个个体的信息收集无关。这个计算没有包括来源于调查过程其他方面的误差估计值。

其他样本设计特点的影响

前面讨论了简单随机抽样误差的计算方法。抽样误差的估计值会受不同抽样方法的影响。如果没有分层,系统抽样产生的误差会与简单随机抽样相同;如果每一层的抽样率是一样的,那么对于同规模的样本来说,分层样本抽样误差比随机抽样样本的要低。

不等抽样率(以不同的抽样率从总体中抽取子群)的设计,是为了增加过度抽样的子群的估计值的精确度。

- 形成总体样本时,一般情况下,对子群的过度抽样会比用简单随机抽样抽取同样规模的样本产生的抽样误差高。
- 如果分层中某些变量变异性较大,那么,用过度抽样导致的那些变量的总体抽样误差会比用简单随机抽样小。

整群抽样产生的误差比同规模的简单随机抽样的误差高,其原因是变量在整群中的同质性比总体的高。同样,最后阶段整群

的规模越大，通常对抽样误差的影响也就越大。

一般情况下很难预料到设计特点对估计值精确度的影响。设计影响会由于研究的不同和同一研究中变量的不同而有所不同。举例来说，假设在各个选定的街区中的每座房屋的建筑类型是相同的，也不知它们是否被哪个人拥有。这样，一个街区的某个应答者报告说自己是一个住宅的主人，那么对这个街区的其他访谈就绝不会获得关于在总体中房屋所有权比率的更多资料。正是这个原因，不管研究者对一个街区只调查一家还是调查 20 家，其估计值的信度都是一样的，基本上与进行访谈的街区的数量是成比例的。另一个极端情况是，某街区里成年人的身高差异与整个城市成年人的身高差异相同。当某街区回应者的同质性与总体相同时，整群抽样不会降低一个给定的身高样本估计值的精确度。因此，为了估计整群抽样对抽样误差的可能影响，研究者应该注意整群或分层所具有的特点以及要得到的样本估计值的种类。

样本设计对抽样误差的影响经常不受重视，经常会看到一些关于置信区间的研究认为，当使用整群抽样时，就可以采用简单随机抽样方式。想要提前知道设计影响的大小是十分困难的。正如所提到的那样，样本设计对抽样误差的影响就每个变量来说都是不同的。当样本设计不是使用简单随机抽样时，如使用整群和分层，这些计算尤其复杂。因为计算抽样误差的能力是运用调查方法的最重要组成部分之一，因此，让统计学家参与具有复杂样本设计的调查，以确保能计算并恰如其分地告知样本误差，是十分重要的。当评估样本误差时，对设计特征的适当考虑被极大地简化了，因为有几个现成的分析程序包将会对此加以调整。

最后，只有把样本设计的特点放在所有调查目标里考虑才能评估出它的合理性。整群抽样可以节约抽样(名单)和资料收集的成本，并且它往往对许多变量的抽样误差影响不大。过度抽样一个或多个组群经常是一个很划算的设计。像在本书提到的很多问题一样，对一个研究者来说，重要的是要意识到所做选择的潜在的成本和效益，并把它放到所有可供选择的设计中，结合调查的主要目的加以衡量。

样本的规模应该多大

对一个调查方法学家而言，在诸多样本设计包括的问题中，最普遍的问题之一是样本的规模应该是多大。在提供答案之前，也许最好先谈谈三种普遍但不是很恰当的确定样本规模的做法。

一个普遍的误解认为，样本的适宜性在很大程度上依赖于包含在样本中的总体的分数——或者是总体的 1%、5%，或者其他比例，这个比例决定着样本的可靠性。已有关于抽样误差的估计值的讨论，没有考虑到包含在样本中的总体的分数问题。前面的公式和表 2.5 中得到的抽样误差估计值，可以通过乘以$(1-f)$使其降低，其中：f=样本中所包含的总体的分数。

当研究者对总体的 10%或更大比例进行抽样时，这个调整对抽样误差的估计值有明显影响。然而，大多数样本即使再大也只包括了总体中很小的一部分，在这种情况下，样本中包含的总体的分数即使增加一点，也不会影响研究者从样本推断总体的能力。

也该注意到与这个原则相反的方面。总体的规模对于从中选出的一定规模的样本能否很好地描述总体并没有实质性影响，假设样本设计和抽样过程的所有方面都是一样的，一个由 150 人组成的样本对一个由 15 000 或 1 500 万人组成的总体进行描述，可以得出同样有效的精确结论。与总的样本规模和其他设计特点相比（例如整群抽样），抽取的总体的分数对抽样误差的影响微不足道，在决定样本规模的时候，对其进行着重考虑是不必要的。

第二种不恰当的确定样本规模的做法比较容易理解一些。有些人受到了所谓的“标准调查研究”的影响，他们据此引申出了典型的或恰当的样本规模。于是，有的人会说全国性调查样本一般来说是 1 500 人，或者说一个好的社区调查样本是 500 人。当然，参考有实力的调查者所认为的对特定总体来说最适宜的样本规模不失是一种聪明的做法。但是，像大多数其他的设计一样，样本规模的确定必须是建立在具体问题具体对待的基础上，要以研究者所要达到的各种研究目标以及对研究设计的许多其他方面的考虑

为转移。

第三种，也是特别需要提到的一种确定样本规模的错误方法，因为它可以在很多统计学的教科书上找到。方法是这样的：研究者先确定能允许的误差极限或所要求的估计值的精确度，一旦明确了所要求的精确度，就可以简单地使用诸如表 2.5 或各种以上所提到的方法来计算能保证这个精确度的样本的规模。

从某种理论角度看，这种方法没有错误。但是，在实践中，它对打算进行真正研究的大多数研究者来说是没有什么帮助的。首先，通常不会仅仅根据估计值的精确度的需要来确定样本规模。大多数调查研究设计是要得出一些估计值，而这些估计值所需的精确度可能是不一样的。另外，让研究者以相当概括的方式来指定所需的精确度是不常见的。除非是特例，一般情况下不能提前指定一个可接受的误差极限。即使在特例的情况下，所适用的方法也意味着抽样误差是调查评估中惟一的或主要的误差来源。当指定样本调查的精确度时，往往会忽略抽样之外的其他方面的误差来源。这种情况下，单以抽样误差为基础的精确度的计算是不现实的、过度简单化的。而且，在研究资源一定的情况下，减少用于回应率、问题设计或资料收集质量的资源去增加样本的规模，反而会降低精确度。

与样本规模有关的抽样误差的估计值在分析所需样本规模中的确扮演着一定的角色，但是，这个角色是很复杂的。

决定样本规模的首要条件是一个分析规划，这个分析规划的关键部分通常不是有关整体样本的置信区间的估计值，而是总体内结合部分总体的一些评估值形成的那些子群的大概情况。典型的情况是，设计过程很快就进行到在需要数据的总体中分出更小的组群，然后，研究者对所需样本规模进行估算以便确定一个小而充分的子群样本。大多数样本规模的确定并不关注对总体的估算，而是关注那些重要的最小子群能容许的最小样本规模。

再回到表 2.5，不看样本规模的连续体的上限而是看其下限。50 个研究对象够不够？如果有人研究表 2.5，就可以发现在样本规模达到 150 到 200 之间时其精确度得到了稳定提高。过了这个点，增加样本规模所获得的效果就会小得多了。

像许多与研究设计有关的决策一样，很少有关于样本规模的研究。有许多方法可以增加调查估计值的信度，增加样本规模就是方法之一。纵然不能说只有惟一正确的答案，然而也能说出三种决定样本规模的不适宜的方法。指定一个总体的分数作为样本规模的方法决不是确定样本规模的恰当方法。抽样误差从根本上取决于样本规模，而不是取决于总体体现于样本中的比例。说一个特定的样本规模是研究一个总体的通常的或典型的方法，实际上也总是不正确的。致力于研究目标的分析规划才是关键性的第一步。最后，用所期望的某个总体变量的置信区间来确定样本规模的大小，是一种极为少用的方法。

作为总体调查误差一部分的抽样误差

抽样过程以三种不同的方式影响到调查估计值的质量：

- 如果样本框排除了一些要描述的人，那么样本估计值就会由于对这些本应包括在内的人的忽略程度的不同而有所偏误。
- 如果抽样过程不是概率性的(probabilistic)，那么样本和被抽样的对象之间的关系就是有问题的。有人可能用其他理由而不是用抽样过程来证明样本的可信度，但是，除非抽样过程让每个人都有被抽取到的机会，否则没有任何统计数字能证明样本对被抽样的总体而言具有代表性。
- 概率样本的规模和设计，以及被评估内容的分布，共同决定了抽样误差的大小。也就是说，偶然变数的产生是由于收集的资料仅仅是关于一个样本的。

通常抽样误差被看作是调查估计值中不可靠性的惟一来源。对于使用大样本的调查来说，产生误差的其他原因可能要更重要一些。本书的主要观点就是非抽样误差应该与抽样误差一样受到重视。当样本设计中包括了整群，或者根本没有形成概率样本时，采用简单随机抽样法也经常出现抽样误差。具有讽刺意味的是，用这些方法得出的抽样误差的估计值会误导读者对样本估计值准

确度或精确度的判断。

如果一个详细的名单被用作样本框,或者采用一个简单的随机或系统抽样方案,或者所有的回应者都是以同样的概率选出来的,那么从一个样本中得出抽样和分析数据可能是十分简单的。就这样的设计而言,表 2.5 和其所依据的公式将会为抽样误差提供一个准确的估计值。但是,甚至用这样简单的设计,当评估调查估计值的精确度时,研究者仍然要考虑到所有的误差来源,包括样本框、无回应和回应误差(这些将在下一章讨论)。此外,当对最好的抽样方法有怀疑,或是采用简单随机抽样有偏差时,一位抽样专家的参与,对设计一个合理的抽样蓝图和正确地分析一个复杂的样本设计的后果都是极其必要的。

练 习

为掌握抽样误差的含义,从同一名单(例如:电话簿)中重复抽取同样规模(不同随机起点)的样本。将有相同特点的(例如:买卖清单)样本按比例放在一起就会产生一个分布图,这个分布图将有一个大约是表 2.5 中已抽取样本规模数一半的标准离差。计算表 2.5 里的几个条目(例如,不同的样本规模和比例),对帮助人们了解这些数字的来源是十分有价值的。



3 无回应:实施样本设计

Nonresponse: Implementing a Sample Design

从被选择的样本中收集资料失败的比例过高是调查误差的主要来源。本章将讨论利用邮件、电话和私人访谈联系回应者并赢得合作的方法。与无回应相关的偏误也将被加以描述，如运用定额抽样以减少因要求获得高回应率所需付出的努力。

概率样本的理念是总体中(或至少样本框中)的每个人都具有一个确定的获得其资料的机会。一个抽样过程会选择一组特定的个体(或某种单元),样本资料的质量取决于那一组中被收集到资料的人所占的比例。在决定样本资料反映总体的程度上,收集资料过程与样本选择过程同样重要。

当然,调查里任何一个估计值的精确性都取决于对某个问题提供答案的人。在每个调查中,有些人虽然愿意成为回应者,但不愿意对每个问题都做出回答。尽管那些针对个体的问题无回应率通常很低,但偶尔也会很高,并且对估计值确实有影响。本章关注的是那些没有提供任何资料的人。

被选作样本但实际上没有提供任何资料的人可分为三类:

- 资料收集过程中没有联系到,因此,没有为他们提供回答问题的机会的人。
- 拒绝提供资料的人。
- 不能按要求提供资料的人(例如:那些病情严重而不能接受访谈的人,使用与研究者不同语言的人,或者因读写技能有碍而不能填写自填问卷的人)。

研究者确定使用的研究步骤极大地影响着样本中实际提供资料的个体的比例(即回应率),也影响着无回应者带给样本资料误差的程度。值得注意的是,与第 2 章讨论过的抽样误差不同,无回应可能使样本“偏见化”,也就是说,使样本系统性地不同于它们所由抽样的总体。在本章中将讨论无回应对调查估计值的影响和减少无回应的措施。

计算回应率

回应率是评估收集的资料有效性的基本参数,即接受访谈(或回应者)的人数除以样本的人数(或单元数)。分母包括了从总体中抽取的所有人,即包括那些虽被抽中但因拒绝、语言问题、疾病或者缺乏易得性而没有回应的人。

有时样本设计会使用删选的方法从总体中选择研究对象。不

包括在研究总体里的被删除的单元不能纳入回应率的计算。所以，空房子、无用的电话号码以及没有符合条件的人居住的家庭（例如，当对老人进行抽样时，某些家庭中没有老人），在计算回应率时就会被排除在外。有一些单元，如果无法获得一些确定其资格的信息，那么，回应率是不确定的。在这种情况下，最好的方法就是通过对没有被删除的单元合格率做出一个保守而又大胆的假设来计算回应率，并且与最佳估计一起报告出回应率的值域。

回应率通常被认为是样本中能收集到资料的人所占样本的百分比，有时更进一步的计算就是样本中能够体现总体的分数。如果抽样框没有排除研究总体里的任何人，那么回应率就等于样本代表总体的百分数。但是，假如总体中只有 95％的人有电话，那么，样本代表总体的百分数的最精确评估值应该是电话调查回应率的 0.95 倍。

弄清计算回应率的详细方法是十分重要的，因为计算方法的不同会导致比较上的困难。例如，某些机构报告了一个“完成率”，即访谈中“接触到的”家庭的比例；由于存在着能抽取却无法接触的单元，因此这个比例总是高于前面提到的回应率。由美国民意研究协会（AAPOR）出版的一本书对回应率的计算和报告方式做了极其详尽的讨论（AAPOR，2000）。

与无回应有关的偏误

无回应对调查估计值的影响取决于无回应的比例及其导致偏误，即与总体具有系统性不同的程度。如果大多数被抽样的人都提供了资料，那么即使没有回应的人是独具特色的，样本估计值仍然是非常准确的。例如，美国人口普查局在进行全国卫生访谈调查时，其成功之处在于完成了对近 95％的所选定家庭的访谈工作，很容易看出，即使那没有回应的 5％的人独具特色，样本得出的结论与总体仍然非常相似。

人口普查局的经验表明：一方面，调查具有极其明确的方向性，另一方面，有时会看到只有 5％到 20％被抽取到的人对邮件调

查做了回应的报告。在这个例子中，最终的样本与最初的抽样过程间几乎没有关系，那些回应实际上是他们的自我选择，因此这样的过程不可能提供任何关于总体特点的可靠的统计数据。

大多数调查研究项目处于这样的两个极端之间。通常，在农村地区的回应率比在中心城市的高。从一个家庭中有责任心的成年人那里获得资料比对预先选定的特定调查对象进行访谈来得容易。一些主题(例如:健康)可能比其他主题(例如:经济行为或公众意见)能引起更多人的兴趣。另外，不同调查研究机构为提高回应率所投入的时间和金钱的差别很大。

并不存在可接受的最小回应率的统一标准。联邦政府管理和预算办公室评价那些与政府相关的调查活动，通常要求回应率超过75%。在美国，学术调查机构通常能在一般家庭样本中被选定的成人那里获得75%的回应率。市中心区或随机电话拨号调查的回应率可能比较低——通常是低得多。

在邮件、电话和个人访谈中，与无回应有关的偏误也有所差异。一般的规律是，支持邮件调查的人大多是对调查主题或研究本身特别有兴趣的人，他们比那些没有兴趣的人更有可能寄回问卷。这就意味着低回应率的邮件调查所产生的偏误可能直接与研究目的相关(例如:Filion,1975;Heberlein & Baumgartner,1978;Jobber,1984)。一项针对那些没有寄回关于健康治疗经历问卷而后来又通过电话进行了补充访谈的人的研究表明，与回应者相比，无回应者通常更年轻、更健康、很少使用保健器材并且大多为男性(Fowler,1998)。

另一个邮寄问卷低回应率导致显著性偏误的例子是经常被引用的《文摘》(Literary Digest)杂志1936年的总统选举民意测验。这个民意调查机构预测兰登会获胜，结果富兰克林·罗斯福以压倒多数的选票获胜。据说，这些调查的样本是从电话簿里抽取的，而1936年，在家里安装了电话的共和党(兰登所属的党派)成员很可能比民主党成员多。除此之外，《文摘》杂志使用了邮件调查，无回应率也是预测失败的原因之一，只有少数的人按要求寄回了问卷。这点在邮寄问卷的调查中很典型:那些希望处于劣势的候选人获胜的人，即兰登的支持者，特别希望能表达他们自己的观点

(Bryson,1976;Converse,1987)。

另外一种在邮件调查中经常出现的偏误是,受过良好教育的人会比受过较少教育的人更快地寄回问卷。因此,任何与教育(它又与收入水平有关)相关的变量的邮件调查的评估都可能带有偏误,除非采取措施以便获得高回收率。

与邮件调查相比,调查对象的可得性(availability)对电话调查和当面调查无回应的影响更大。很明显,如果资料收集是在周一到周五的早上 9 点至下午 5 点之间进行,那么得以进行面谈的人将具有一些共同的特点:由于那时在家的往往是一些没有工作的人,因此调查很可能会得出家庭妇女、带孩子的母亲、无业人员和退休人员的比例偏高的结果;这些人一般不忙于义务活动和社会活动;他们更可能是一些很小的孩子的父母;规模大得家庭比只有一两人的小家庭更有可能有人在家。

调查对象的可接近性(accessibility)也会导致无回应产生的偏误。使用个人访谈的国家调查,在市中心区获得的回应率总是比在郊区和农村地区的低。导致这种情况产生的原因主要有三点:第一,在市中心区找到单个的个体样本是十分困难的;第二,在市中心区,越来越多的人住在高层公寓中,使得研究者很难直接与那些人接触;第三,市中心区有很多地方不适合访谈员晚上访谈,因此,他们不能判断出那些不容易找到的人什么时候恰好在家。

在过去的 20 年里,市中心区持续增长的人口和其他广泛的社会变革(单亲家庭越来越多,有孩子的家庭越来越少,女性劳动力越来越多)使得越来越难获得高回应率。现在有了一些回应率下降的迹象,有些机构虽然维持了回应率,却额外增加了投入和工作量。格罗夫斯和库珀(Groves & Couper,1998)对无回应的全国性趋势做了完整的总结。

有些材料表明,使用电话调查可以减少市中心区与农村地区间的回应率差距,因为它有可能对城市家庭进行全面覆盖,能够与居住在高安全建筑里的人保持联系,并通过努力可以找到在家里的个体。

从另一方面来说,受教育少的人不是十分愿意接受随机电话拨号的访谈方式,年龄超过 65 岁的老人亦是如此。这种偏误很少

出现在个人访谈调查中。格罗夫斯与凯恩(Groves & Kahn ,1979),坎内尔、格罗夫斯、马格利、马斯维茨和米勒(Cannell, Groves, Magilavy, Mathiowetz, & Miller,1987),格罗夫斯与利伯格(Groves & Lyberg,1988)以及和格罗夫斯与库珀(Groves & Couper,1999)提供了电话调查无回应偏误的良好评价。

最后,偏误还与那些不能接受访谈或不能填写表格的人相关。这些人通常只占总体的一小部分。不考虑那些住在医院的人,对于保健利用或开支的调查来说会是一个非常严重的疏忽;排除了街区中那些不会说英语的人,也会对调查结果产生明显的影响。如果没有采取专门的措施去从特定的组群中收集资料,样本估计值就会只适用于更有限的总体:对给定的资料收集措施来说,这个总体实际是那些有机会回答问题或提供资料的人构成的总体。

虽然在访谈调查中的回应者与无回应者存在着人口统计学上的差异,特别是随机电话拨号基础上的电话调查,无回应者对调查估计值的影响不是十分明确。基特、米勒、科弗特、格罗夫斯和普雷瑟(Keeter,Miller, Kohut ,Groves, & Presser,2000)提供了对两个电话调查结果的详细比较:一个调查的回应率为 36%,另一个为 60%。调查包括了人们对政治和社会的态度,但是,从两个调查的估计值来看,很少有统计数据上的显著的不同。

关键的问题是无回应与调查主题相关的程度。回应者对研究主题的兴趣可能对邮寄问卷调查的回应率起着重要的作用,比它对访谈获得合作的影响更大。这看起来似乎是很有道理的,并且与现有的证据相一致。然而,无论如何,我们对什么时候、有多少无回应会或不会影响调查估计值缺乏可靠的预测信息。

表 3.1 提供了一个例子帮助我们了解无回应对调查结果的潜在影响。假设某样本有 100 个人,但只有 90 个做出回应的人(回应率为 90%)。在这 90 人中,有 45 人说“是”,其他的 45 人说“否”。还有 10 个人(无回应者)的态度我们不知道,假如他们都说“是”,那么事实上就有 55%的人说“是”。因此,就有:

- 回应率=90%
- 50%的回应者说“是”

表 3.1 关于由无回应引起的潜在误差的例子

	回 应* (知道他们的答案)**	无回应 (不知道他们的答案)	总的样本 (假如每个人都回答时的答案的可能范围)
"是"	45	0～10	45～55
"不是"	45	0～10	45～55
总数	90	10	100

注：* 假设回应率是 90%

* * 假设有 50%的回应者说"是"

在最初选定的样本中，回答"是"的人的实际比例最多不超过 55%，最少不低于 45%。例子中可能出现的范围写在表 3.1 的右边那一栏。

表 3.2 列出了关于回应率范围的逻辑，从中可以看出，当回应率低的时候，由无回应引起的潜在误差十分大。

表 3.2 在某回应率水平上，样本里有 50%的人回答问题时，真值的可能范围

前 提	回应率				
	90%	70%	50%	30%	10%
假设回应者中有 50%的人做出了某一项回答，再假设样本中每个人也做了应答，那么真值的范围应是：	45%～55%	35%～65%	25%～75%	15%～85%	5%～95%

资料来源：根据杰克·埃利森(Jack Elison)和米切尔·D·埃利森(Mitchell D. Elison)所制的有关个人交流表改编。

将无回应偏误的影响因素加以对比可能具有启发性，例如，对表 3.2 与前一章(见表 2.5)中提到的抽样误差的影响因素进行比较分析。人们通常不知道无回应怎样产生偏误，但是认为无回应不产生偏误是不对的。调查人员要努力确保回应率达到一个合理的水平，并且避免在回应者和无回应者之间采用导致系统性差别的措施。这是保证调查估计值精确性的重要方法。

减少电话或当面访谈调查中的无回应率

就像有人经常说依据量多的样本比量少的样本所得出的结论更可靠一样,在其他方面等同的情况下,也可以说与具有较高无回应率的调查相比,具有较高回应率的调查会产生一个更好、偏误更少的样本。无论做出什么设计决定,研究者必须考虑对减少无回应率做出多大努力。

在电话和个人访谈调查中为了得到较高的回应率,必须处理好两个不同的问题:与选定的个人进行联系和赢得他们的合作。

为了减少由于缺乏联系不上而产生的无回应率:

- 集中在晚上和周末多打电话,打电话的次数取决于当时的情况。在城区,每家至少要打 6 次电话。对于电话调查而言,如果越多地通过电话实现调查目的,总费用就会越低。许多研究机构把 10 次作为最低限度,而实际操作时大都多于 10 次。
- 为访谈者提供弹性进度表,能使他们在回应者方便时随时进行访谈。

为了赢得合作:

- 如果可能,先寄一封预告函。这可以打消一些回应者的顾虑,并且访谈员也可能感到更有信心。
- 提供关于研究项目目的的精确信息,让回应者明白他们的帮助是十分重要和非常有用的。
- 要确保调查任务和使用的措施不会对回应者产生不利影响。
- 找到得力的访谈员,让他们明白回应率的重要性。尽快找出不利于提高回应率的访谈员,重新培训或停用那些办事不力的访谈员。

维持采访者坚持采访的权利与被采访者拒绝回答问题的权利之间的平衡是不容易的。调查中的参与应是自愿的,勉强参与的人,经常是由于对调查的无知。格罗夫斯和库珀(Groves & Cou-

per，1998)对回应者勉强参与访谈的原因进行了建设性的分析。

最好是在回应者拒绝之前告知他们调查的目的。主流观点认为，大多数的调查都具有有价值的目的，如果有人拒绝回应，作为结果的资料也将不太准确。访谈员应该努力让被访谈者明白他们正被要求做什么以及为什么这样做。在被调查者刚开始拒绝后再次提出让他们重新考虑也是十分有效的做法。有相当一部分拒绝回应者是由于在不方便的时候被访谈，而不是他们不愿意接受访谈。在那些最初拒绝回答的人中有 1/4 到 1/3 的人后来又同意接受访谈。

对回应者来说，访谈过程通常是一次积极的体验。如果调查研究项目是由负责任的组织进行的，回应者的回答就会得利严格保密。通常，调查研究者会像精神病学家和记者保护他们的信息来源一样遵守保密原则。很多调查研究项目有合理动机，这使得大多数人都愿意提供帮助。如果访谈员愿意在回应者方便的时候进行访谈，对大多数回应者而言，将没有时间上的特别压力。最后，大多数回应者报告说被访谈是一件愉快的事。人们喜欢有机会向一个很好的倾听者谈论自己。

随机电话拨号对取得高回应率来说是一个有特别意义的挑战。当使用名单辅助法选择样本时，就可把信寄到已知地址的家庭。但是，就算这样做了，保证人们接电话回答问题仍然是研究者的一个难题。研究者在电话里开头所说的几句话是成功的关键。由于电话常被用作一些推销和募捐的手段，因此人们听到是陌生人打的电话时都会产生抵触情绪。在此情况下，有时即使再真诚的机构获得的回应率也只有 40%到 60%，特别是当调查的主题不能立刻吸引人时回应率更低。获得高回应率的困难在于随机电话拨号作为一种抽样方法的内在局限性。

最后，应该注意，事实表明，在访谈以前付给被访谈者费用会增加私人访谈和电话访谈的回应率。从历史上看，报酬最初是在要求回应者回答不同寻常的问题的访谈研究中使用的，例如，多次性的访谈。但是，最近的研究表明，在一次性的访谈里，报酬也会对回应率有所影响(Groves & Couper，1998；Singer，Van Hoewyk，Gebler et al.，1991；Singer，Van Hoewyk & Maher，2000)。

减少邮件调查中的无回应率

减少邮件调查中无回应率的问题稍微不同。与回应者取得联系是电话和个人调查中十分重要的一步，而在邮件调查中，只要研究者有一个准确的邮件地址就行了，大多数人迟早都会回家取邮件。然而，最主要的困难是如何在没有访谈员参与的情况下引导回应者完成回应工作。

写信说服大多数人做某事并不是一种很有效的方法，个人联系比写信更有效。现在有大量关于使邮件联系更有效的设计策略文献。是用彩纸还是白纸印刷调查问卷？怎样才能使信笺抬头给人留下深刻印象？怎样注释？应该支付回应者什么报酬？是在寄信的同时寄现金，还是承诺一回收到问卷就付报酬？回应者的信件是应该用蓝墨水写吗？贴邮票比已付邮资的信封好吗？

总之，能让使用邮件的调查问卷看上去更专业、更人性化或者更吸引人的所有措施都会对回应率产生正面影响。从总体上来关注这些细节很值得，迪尔曼(Dillman，2000)详细评述了这些问题。另外，使调查问卷容易完成可能也是十分重要的。有关的设计细节将在第6章讲述，但这里需要指出以下三点：

- 层次要清晰，以便更容易看出其进展的步骤。
- 问题要被分成一个个的小问题，容易阅读且不凌乱。
- 回应工作应该很容易做。除非他们自己选择，一般不要叫回应者提供书面答案。在回答问题时，只需他们勾复选框，圈数字或者其他简单的方式。

迪尔曼(Dillman，2000)还提供了更广泛的有关自填问卷调查设计的指导。

几个已发表的研究指出，提前向回应者支付报酬会提高回应率(Fox, Crask, & Kim, 1988; James & Bolstein, 1990)。几乎所有的调查都表明支付报酬的好处，即使是医生所进行的调查也是如此(Berry & Kanouse, 1987)。但只有提前付钱才能对回应率有所影响。延迟付款、提议将报酬捐赠给慈善团体和先登记后支取

都不能增加回应率(Warriner, Goyder, Jersten, Hohner, & McSpurren, 1996)。最后,毫无疑问,好的与差的邮件调查的最大区别是研究者与无回应者反复联系的程度。正如迪尔曼(Dillman, 2000)加以概括的,合理的步骤应该包括以下几点:

- 在最初寄出信件的 10 天后,给所有无回应者寄提醒卡(Reminder card),以强调研究和高回应率的重要性。
- 在寄出卡片大约 10 天后,再给剩余的无回应者寄出一封信,并再次强调高回应率的重要性并附上问卷,以防有些人已经将第一次寄去的问卷扔掉了。
- 如果回应率仍不让人满意,也许下一步最好就是给无回应者打电话。如果没有电话号码或者打电话的花费太大,追加邮寄劝说函、发送夜间电报、通宵递送或其他补救措施都将是有所助益的。

让回应率达到合理水平的困难在于样本的性质、研究的特点、人们的积极性如何以及人们完成问卷的难易程度。如果样本是由有积极性的受过良好教育的个体组成,那么,想要得到较高的回应率将会更容易。迪尔曼只使用邮件,就在一般的总体样本中得到了 70% 的回应率(Dillman, Carpenter, Christensen, & Brooks, 1974)。当然,如果再给无回应者打电话,回应率将会更高。如果研究者坚持努力,保证研究是合理地精心构思和仔细设计的,邮件调查可以得到可接受的回应率。

如果研究者打算再次和无回应者取得联系,就必须知道哪些人没有寄回问卷。这个过程不需要拐弯抹角或者复杂化,在问卷或者寄回的信封上写上简单的编号即可,最好是在信里告诉人们该数字的用途。

有时,研究者可能想给回应者担保是不能被识别的,这时有两种可供选择的方法,即更仔细的工作和采用补救措施。寄给回应者的问卷上没有识别代码,但是在问卷上另附一个单独的有回应者代码的明信片,其内容如下:

> 敬爱的研究员,我在寄回问卷的同时寄回了这张明信片。鉴于我的问卷完全是匿名的,但是这张明信片会

告诉你不用给我再寄索回问卷的提醒卡了。

这种方法保证了回应者的匿名性,同时又告诉研究人员某人完成了问卷。有人也许会认为回应者会只寄回这张明信片以避免再收到提醒信,但是这种事情很少发生,事实上,寄回的明信片数量和问卷数量大多数情况下几乎是一样的。希伯(Sieber,1992)探讨了一些收集匿名资料的方法。

减少互联网调查中的无回应率

尽管互联网调查是一个相对较新的方法,没有像邮件和访谈调查那样的经验性的文章,但是其难度和挑战可能接近于邮件调查和访谈调查。问题仍在于要在没有访谈员的情况下促使人们做出回应。有积极性和对研究感兴趣的人最有可能做出回应。使作业简单易行、反复联系、使用多种方法与回应者相联系以及为那些开始时没有回应的人提供可选择的回应方式,这些可能都是最大限度地提高回应率的关键。像邮寄问卷调查一样,调查内容与被调查总体间的一致性以及调查样本在互联网使用者中的易得程度,可能都是保证获得高回应率的重要因素。迪尔曼(Dillman,2000)对这些问题做了详细地概括。

多种形式资料收集

要最大限度地减少调查无回应的最好方法之一就是使用多种方式相结合的方法收集资料。要注意的关键问题是如何接近调查对象、提高被调查人的积极性和降低成本。多种方式结合的方法可以让研究者接触到那些用一种方式很难找到的人,同时也能从本来没有多大内在积极性的样本成员那里收集到资料。例如,一个吸引人的方案是使用电子邮件或邮寄问卷作为资料收集的第一步,接着与无回应者进行电话访谈。另一种很有效的设计,是把电

话和面对面的访谈结合起来。

多模式调查中的一个关键问题是各模式所获得的资料间的可比性。有的问题的答案会受到资料收集方式的影响,有的则不会。为了将通过不同方式收集到的资料整合起来,资料具有可比性是十分重要的。在第4章中将就这些问题进行更多的探讨。

修正无回应率

有些无回应是不可避免的,这种情况下,有三种尽可能将无回应导致的误差降低的方法:使用代理回应者(proxy respondent)、进行统计资料的调整以及对无回应者进行重新调查。

代理回应者

很多调查一般是从某个家庭的回应者那里收集其他家庭成员的资料。如果一位回应者不能或不愿接受访谈,一种选择就是让该家庭的另一个成员来代替那人回答。对代理回应得到的资料的质量的大多数研究表明,对于大多数话题,代理回答不如回应者自己回答的效果好。此外,只有很少的研究者会接受代理回应者对主体态度诸如感受、知识或意见的报告。但是,对于事实性的资料,如果有一个现成的知识渊博的代理者,那么使用代理回应者也能够有效减少无回应导致的误差。格罗夫斯(Groves,1989)评论了那些论述代理回应者所提供资料质量的缺乏说服性的文献。

统计资料调整

由无回应所引起的统计资料的调整工作如下:假设某总体中有20%的成人是65岁或65岁以上的老人,由于各种无回应的存在,样本中却只有10%的回应者在65岁或65岁以上。在调查完成以后,研究者可以给那些在65岁或65岁以上的回应者的答案加权,以便使他们与20%的回应者等效。鉴于65岁或65岁以上的人的答案与年轻人的答案的不同,这样所得的最终估计值可能会更好。但是,这个结果需要假定65岁或65岁以上的回应者和该年

龄段的无回应者是相似的,这却可能是不符合实际的。

很多机构是通过统计处理来降低由无回应引起的估计值的误差的。卡滕(Kalton, 1983)、利特尔和鲁宾(Little & Rubin, 1987)、格罗夫斯(Groves ,1989)仔细探讨了这种调整的技巧、设想和局限性。毫无疑问,这种调整会提高某些估计值的精确度。即使如此,大多数人都会认为,减小调查估计值中由无回应导致的误差的可靠方法首要的还是降低无回应率。

调查无回应者

假设在一次邮寄问卷调查中,被抽样的人中有 60%做了回应。如果研究者认为对无回应者进行电话或个人访谈其中很多人就能回答问题,但是对所有无回应者采取这种措施缺乏资金,那么可以选择对需要更高联系费用的无回应者抽取一个样本的办法进行。

对无回应者进行调查的工作有两种不同的效用:第一,第二轮收集到的资料可以用来评估最初样本里的偏误方向和数量。当然,第二轮资料收集中也会出现无回应率,因此它的资料不能完全代表所有的无回应者。但是,考虑到这个局限,这些资料仍可用于统计调整。第二,如果第二轮的资料收集中所用的问题重复初次调查中的问题,那么它的资料就可以补充到最初所收集到的资料上。在追加资料的过程中要注意对追加的资料进行加权处理,以适应只对无回应者的一个样本进行了补救的事实。

如果对一半的无回应者做了补救调查,在与前次收集的资料整合时,这一阶段中的回应者也应该进行系数为 2 的加权处理。另外,这个结果也可以很好地反映调整后的样本总体的百分比(一个调整后的回应率),计算公式如下:

$$\text{调整后的回应率}=\frac{\text{最初的回应率}+2\times\begin{array}{c}\text{从无回应样本中}\\\text{得到的回应率}\end{array}}{\text{最初符合条件的整个样本}}$$

非概率(或可变概率)样本

(nonprobability/modified probabilily samples)

本章前面所提到的都是按照概率样本的假设进行设计的,因此,回应者是通过客观程序选定的。研究者面对的问题是去收集那些被选定对象的资料。但是,有一种普遍使用的方法(或一系列方法)是不选定特定个体,并且也无法计算出他们的回应率。之所以要讨论这些非概率样本,实质上是因为它们的回应率不可计算,并且它们与那些回应率很低的样本相似,由此导致使用者可能不知道他们所使用的资料的局限。

调查研究者之间关于概率抽样的重要性的争论是很有趣的。联邦政府通常不会为那些不是建立在概率抽样技术基础上的对总体特点进行评估的调查研究提供经费。大多数学术性调查机构和很多非赢利性调查机构都有类似的抽样方式。同时,大多数较大的民意测验团体、政治民意测验团体和市场研究机构主要使用非概率抽样方法(参见 Converse,1987,对这种区别的历史根源进行了探讨)。

概率抽样的基础是样本里的人是在预定的程序中获得的,这个程序为被定义的总体成员设定了一个选择率。除了这些,回应者的特点或研究人员的处置权都不会影响到某人成为样本的可能性。虽然,可变的非概率抽样的具体措施是变化的,但是,在最后阶段,他们都受调查员的处置权和回应者的特点的内在属性的影响,这些虽不是样本设计的一部分,却可能会影响成为样本的可能性。接下来将讲到其中两个最常用的步骤。

个人访谈研究会包括非概率性的抽样,因此,在非概率性抽样中研究者抽样街区的方式与抽样街区形成区域概率样本的方式很相似。其区别在于,一旦某个街区被抽中,访谈员就走访那个街区并找到居住在那个街区的人以来完成特定数量的访谈工作。没有那个街区的特定家庭单元名单,访谈员随意访问那个街区的住户,对不在家的人员也不进行回访。

电话调查使用了相似的方法。选择一个特定的电话局,或者一个电话局里的一组号码,其目的都是为了完成一定数量的访谈。如果在电话访谈时没得到答案或者没有回应者,就在同一组中找出另一个号码进行电话访谈,直到得到足够的访谈数为止。如果抽样的第一阶段按照总体的地理区域分布方式确定样本的分布状况,那么在选取家庭和回应者时,会产生三种十分明显的偏误。

在个人访谈而不是电话访谈中,访谈员可以自由选择去哪家访谈。事实证明访谈员更喜欢去那些富有吸引力的住户,更愿意去那些住在一楼的家庭,而不是二楼或者三楼的,愿意去那些不养狗的家庭。影响访谈员选择访谈对象的其他一些因素,留待读者自己去想象。

一些研究机构试图通过指定从哪里开始以及规定不得略过哪些家庭单元的办法,来限制访谈员的处置权。但是,由于预先没有该街区抽样单元的名单,事实上不可能判断访谈员是否执行了这些指令。另外,如果想预先有该街区抽样单元的名单,那些本来用这种方法可以节约的成本就会被抵消掉。

非概率方法的第二个偏误特点是可得性的干扰。假如访谈员不对那些先前没人在家的家庭进行回访,那么有人在的家庭被选中的机会就比那些平时没人在家的家庭高。性别、年龄、工作状况、家庭规模和有否孩子等问题都与可得性有关。

这种无法控制的抽样方式会产生明显的样本偏误。提高样本质量的最常用的方法是对明显偏误引入配额。例如,要求访谈员对特定街区或电话整群中的一半男性和一半女性进行访谈。偶尔,也会设定一些额外的限制条件,例如预期的种族构成、成年人中老人和年轻人的数量。重要的是不给配额制定太多的强制条件,否则访谈员会用很多时间打电话或在街区里徘徊以寻找符合条件的回应者。

配额或非概率样本的最后一个固有偏误与赢得合作有关。如果回应者说自己很忙或不方便接受访谈时,就说明访谈员没有激励起对方的合作。如果对项目没有进行有效的介绍,绝大部分人不会对提供帮助感兴趣。没有努力向人们介绍的研究项目很容易导致他们拒绝合作,这种情况下,那些忙于工作的人、对研究或对

研究的特定主题不了解或不感兴趣的人，都会导致样本偏误。

萨德曼(Sudman，1967，1976)认为，包括那些努力与无回应者进行联系的调查在内，所有的调查中都会存在无回应的情况。一旦知道某个人不愿意合作或打了几个电话都无法联系到，萨德曼建议从同样街区再找一个代替者，这事实上会提高估计值的准确率。他认为找属于样本中的邻居做替代比找非样本邻居做替代更好。当对访谈员的处置权施行控制时，如萨德曼所说，从减少回访中节省的开支在很大程度上被增加的管理成本抵消了。

非概率抽样方法可以为个人访谈调查节约成本(少于电话调查)。另外，从它们的可比较性方面看，其结果看上去常与概率抽样所得的资料类似。即使如此，有两个事实值得注意：第一，由于节约成本的关键在于避免回访，由此导致在大多数非概率样本民意测验中，总体中只有 1/3 的人有机会被抽中(即第一次电话联系就在家的人)。一个只给总体中 1/3 的个体被抽取机会的样本，也就是具有已知不同特点的 1/3 的总体，很可能在影响样本统计数据的很多方面不具有典型性。鲁宾逊(Robinson，1989)提供了一个被认为做得很好的非概率样本却歪曲了事实的例子。他比较了两个调查的结果，一个调查是采用概率样本，另一个是采用非概率样本，它们的目标都是对艺术兴趣和出席与艺术有关的活动的评估。非概率样本调查大大高估了人们对艺术的兴趣。第二，概率理论和抽样误差的假设常被吹捧为可以描述非概率样本的信度，其实它们并不适用。如果有代替者存在，样本就不是一个概率样本，尽管它有可能以合理的实际方式分布于总体中。

有时非概率样本也是有用的。亨利(Henry，1990)描述了各种各样的非概率样本以及它们的适用条件。然而，如果研究者决定使用非概率样本，他应该告诉读者样本是怎样被抽取的，样本在易得性上和自发接受采访的过程中可能产生偏误，以及计算抽样误差的一般假设在这个样本中的不适用性。这样对读者的提醒目前还不普遍。在很多情况下，非概率样本严重不具有代表性，因此对社会科学研究的信度构成了严重威胁。

无回应是误差来源之一

无回应是一个复杂的、重要的调查误差的来源。表 3.2 表明了无回应对研究结果的重大影响。虽然我们可以计算回应率,却不知道无回应对资料的影响程度。基特(Keeter ,2000)等的研究表明,一个具有相对的低回应率的调查结果与一个有较高回应率的调查结果非常相似。

关键问题是我们常缺乏关于无回应何时可能或不可能产生关于调查内容偏误的准确资料。由于缺乏这些资料,所以对最大限度地提高回应率的争论主要是为了获得信度。回应率高的时候,由无回应引起的潜在误差的可能性就小;回应率低的时候,由无回应引起的潜在误差的可能性就大,调查结果的批评者就有有力的证据说资料是没有信度的。

练　习

如果挑选一个家庭样本作为抽样 18 岁以上成年人的第一步,回应率是完成访谈的人数除以样本中被抽样程序选定的参与研究的个体的总数。你在计算回应率时会把以下各组包括在分母里还是排除在分母外?为什么?

- 无人居住的家庭
- 出去度假的人
- 临时住院的人
- 拒绝访谈的人
- 居住者都小于 18 岁的家庭单元
- 不会讲研究者所使用的语言的人
- 家庭其他成员说他有精神病或者神志不清不能参与访谈的人
- 在访谈员给他打电话时一般都不在家的人
- 那些去上大学的人

确定出你的样本统计数据(和你的回应率)会适用的总体。



4 资料收集的方法

Methods of Data Collection

资料收集方法的选择，诸如选择邮寄问卷、互联网、电话、个人访谈还是团体填答方法，与样本框、研究主题、样本特点以及可拥有的人员和设备有直接关系；它与回应率、问题形式和调查成本也有关系。电脑可以用于所有的资料收集方法。本章将论述有关资料收集方法选择的注意事项以及不同资料收集方法选择导致的不同结果。

作为一个研究者，在收集资料的过程中，必须对资料收集的方法做长远的考虑。在调查中，是让访谈员进行提问和记录，还是让调查对象独立完成？如果在资料收集的过程中使用了访谈员，就需进一步考虑访谈员是该亲自去还是通过电话访谈。如果回应者在没有访谈员的情况下自己阅读和回答问题，那么，就存在怎样向其提供问卷的问题。有时调查问卷是以团体或个人的方式发放并当场回收。在家庭调查中，以分发或邮寄的方式发放问卷，并以同样的方式回收。如果调查是在互联网上进行，就可以通过电子邮件或者要求回应者进入某个网站回答问题的方式进行。

尽管大多调查只使用一种资料收集方法，但是，使用多种方法结合的方式也是十分普遍的。例如，有时个人访谈中存在是让回应者独立填写问卷，还是直接把答案输入便携式电脑中等一系列问题。计算机收集资料的方式也已经与电话访谈相结合。为了减少无回应率，对没有寄回邮件问卷的那些人，则由访谈员打电话或亲自登门与其联系；对于那些没有电子邮件地址或没有回应的人就可以用邮件的方式作为补充；对于那些访谈员无法在其家庭中找到或者已经搬家的调查对象，则可以通过电话或自填问卷的方式进行调查；在一些家庭调查中，对可以得到电话号码的家庭可以通过电话访谈，而对找不到电话号码的家庭则使用个人访谈。

各种资料收集的方法都可以与电脑相结合。在进行电话访谈和亲自上门访谈时，访谈员一般都会用到电脑。更有意思的变化是让回应者把回答直接输入电脑。网上调查是完全没有访谈员介入的。还可以通过让回应者在电脑上回答问题的方式，在其所在的地方——诸如医生办公室等——收集资料。最后，可以通过电话的电脑语音提示系统提出问题，并由电话的数字按钮来选择答案。

在调查中，会出现使用某种方式进行资料收集所取得的效果最好的情况。本章的目的，就是讨论如何针对不同情况选择恰当的资料收集方式。

资料收集方法选择策略中的主要问题

抽 样

研究者计划获取样本的途径是与最好的资料收集方法相联系的。不同的取样方式会影响资料收集的难易程度。如果是从一堆名单中抽样，那么名单上的信息就显得十分重要。如果名单上没有详细的通讯地址、准确的电子邮件地址或正确的电话号码，想以上面所提到的方式收集资料，就十分困难了。随机电话拨号的出现极大提高了电话收集资料的可能性，因为它使每个有电话的家庭都有被选中的机会。如果把没有电话的家庭排除在外，也许这就是获取具有一般性家庭样本的最节省的方式。

当然，也可以使用随机电话拨号方法简单抽样并与家庭进行最初的联系，然后通过其他方法收集资料。一旦联系到某个家庭，研究者就应该向他们询问地址并获得邮寄问卷或派访谈员走访的许可。这种方案十分适合于寻找一个难得的总体，因为对同一任务而言，通过电话进行抽样和筛选(the screening via telephone)比个人访谈更节约成本。这些方法的困难之处在于资料收集过程中双方的合作情况。

当获得样本的依据是一组地址时，这些地址既有可能是从名单中获得的，也有可能是从区域概率样本中得到的，那么电话、个人访谈和寄邮件的方式都是可行的。当然，如果有详细的地址，就可以采用派访谈员的方式。另外，还有可能通过商业途径或对姓名和地址进行编排的互联网机构，以及电话查询服务来获取电话号码。

这些方法不会提供所有地址的电话号码。因此，就必须用其他的资料收集方式对其加以补充。使用电话访谈通常是可行的，然而，仍然要对电话访谈的优越性有所认识。如果是从一些详细的地址清单中获得的样本，那么使用邮件的方式也是可行的。在城市里，有大量多单元的公寓。在进行调查时，除了街区地址外，

还要确定样本的公寓单元(或家庭成员姓名),否则,寄出去的信不是被退回就是寄错了地方。在农村,也存在这个问题,但是现在农村的大多数家庭都有了明确的地址。邮件调查的可行性依赖于所提供的样本地址信息的详细程度,并以此来确定选择的家庭。

包含了电子邮箱的样本名单,则为资料的收集开通了另一条通道。虽然对于一般的总体调查来说,使用电子邮件的方法并不合适,但是有很多人(雇员、学生、专业机构的成员)都普遍使用电子邮箱,并且电子邮箱也是很容易获得的。在这种情况下,把网络作为主要的或至少是其中之一的资料收集方法是十分有意义的。

最后一个要讨论的与抽样相关的问题是对回应者的选择。如果样本的形式是个人清单,那么,任何收集方法包括邮寄都是可行的。但是很多调查认为在资料收集时,选择一个明确的回应者是必需的。如果向一个家庭或组织邮寄问卷,研究者就很难控制由谁回答问卷,因此,如果对回应者的选择存在争议,那么访谈员的介入就是不可避免的。

总体的类型

总体的读写能力和合作的动机是选择资料收集方式时要考虑的两个突出的问题。相对于利用访谈员的方法而言,自填问卷就要对回应者的读写能力提出一定的要求,如果考虑利用网络收集资料,那么,对回应者而言,电脑的使用技术则是一个附加的要求。回应者中那些没有受过多少教育,英语的读写水平不高(但会说英语),视力不好,不经常使用电脑,以及易于生病或劳累的人都会认为使用访谈员比自填问卷的方法要更简单易行。

就使用自填问卷的方法而言,存在的另一个问题是回收问卷。如果没有访谈员的介入,回应者的内在动机就对返回问卷的行为起决定性作用。对某研究问题特别有兴趣的人就更有可能寄回问卷(Fowler,1998; Heberlein Baumgartnen,1978;Jobber,1984)。在这种情况下,如果研究人员是从一群读写能力很强,并且对研究很感兴趣的人中收集资料,使用邮寄或者电子邮件的方法就更会吸引人们的注意;相反,如果是从一群读写能力差或对研究的兴趣和动机都很低的人群中收集资料,那么访谈员介入的方式就更为

恰当。另一个要考虑的问题是要便于联系，访谈员介入方法的使用取决于由访谈员与回应者的联系以及对资料收集的安排。自填问卷的一个突出的优点就是只要提供的联系地址是正确的，回应者就会收到问卷；而且，即使回应者很忙，他也会在方便的时候完成问卷。假如调查是与工作相关的，忙于工作的人习惯于以电话的方式进行，并且要预约；通常会有人为他们安排时间表。以预约的方式进行电话访谈是从这些工作忙碌的人那里收集资料的最好方法。

问题形式

一般说来，如果一个人打算完成一个自填问卷，他必须要使自己接受那些封闭型问题，也就是说那些问题只需他通过勾选或圈点由研究者预先提供的一系列答案来作出回答。从某种程度上说，问题的难易程度是回应率的重要影响因素。第二，自填问卷中的开放式回答经常会包含一些无用的信息。由于没有访谈员参与从中找出明确的、与问题目标一致的内容来，因此，回应者之间的答案没有可比性，而且很难对它们进行编码。如果说这样的应答最终有用的话，那也是作为描述性的材料而不是量化资料。

尽管开放性的问题通常需要有访谈员参与，但在有些情况下仍然可以较好地将其处理成封闭型自填问卷的形式进行。一个很好的例证是，在研究者要用相似的方式调查大量问题时，如果让访谈员去读一长串相似的问题会是一件十分笨拙、枯燥的事情。在这种情况下，就应该把那些问题通过问卷或电脑以自填的形式进行，从而以一种令人愉快的方式提高访谈的工作效率。

当问题的回应项很多或很复杂时，自填问卷法就发挥出其优势。在个人访谈中，一般要给回应者一张列有回应项的卡，以帮助他们记住所有的可能答案。而电话访谈则需要通过下述三种方法予以调整：

第一，研究者可以限制电话访谈的回应规模。有人认为，在电话调查中最多 4 个选项是合适的，在多数电话调查中，以 2 或 3 个选项为主。

第二，如果电话访谈员阅读速度很慢，那么可以使用一个较长的问题清单，读两遍，在访谈员阅读的过程中，回应者就可以随即

选出自己认为正确的选项。通过这种方式获得的答案和以发放书面清单的形式而获得的答案是否一致，目前尚未得到证实。就某些种类的问题而言，回应项的阅读顺序会影响到回应者的回答(Bishop, Hippler, Schwartz, & Strack, 1998; Schuman & Presser, 1981)。

第三，研究者可以将一个复杂的问题分成两个或多个更简单的问题。例如，通常需要人们在超过4种或5种选项中报告他们的收入情况，对一个有9种选项的问题，就可以分成如下两段进行：请问你们全家的收入是少于30 000美元呢，还是在30 000美元到60 000美元之间或是在60 000美元以上？然后，根据回答的情况，访谈员要另外问一个有三种选项的问题，诸如：那么，请问你们全家的收入是少于40 000美元呢，还是在40 000到50 000美元之间或是在50 000美元之上？这些变化起到使问题更易回答的作用，但是问题形式本身有时会对答案造成影响(Groves, 1989)。

对那些包括复杂情境或事件的描述和那些需要图片或其他视觉提示的问题形式，就不适合采用电话调查。如果这种变量的确定是调查中的重要组成部分，那么就需要使用电话调查之外的其他调查方式。然而，研究者们已经证明，他们能够使大多数问题适用于电话调查。通常，将访谈规划调整适用自填问卷较之反过来调整自填问卷规划以适应访谈要容易。因此，如果要将一个量表要既运用于访谈也用于自填问卷，那么先按照访谈方式设计是明智的。

问题内容

很多研究把不同资料收集方式所得到的结果进行了比较(例如，Cannell et al., 1987; Groves & Kahn, 1979; Hochstim, 1967; Mangione, Hingson, & Barret, 1982)。德·路乌与范·德·邹文(De Leeuw & Van der Zouwen, 1998)以及迪尔曼(Dillman, 2000)对结果进行了细致的总结。对于大多数被研究过的调查问题来说，由个人访谈、电话访谈和自填问卷解决的问题总数分配是很相似的。

研究者各有依据地认为，在处理一些敏感话题时，使用这种或

那种方法会更有优越性。自填问卷法被认为是其中最好的方法，因为回应者不必直接向访谈者承认反社会或相反价值的性格或行为。另外一些人认为，电话访谈不涉及个人的空间，这将有助于人们报告一些负面的事件或行为。另外，随机电话拨号提供了几乎完全匿名的调查过程，因为研究者没有必要知道回应者的姓名和家庭住址。还有一些人则认为，个人访谈是询问敏感性问题的最好办法，因为访谈员有机会和回应者建立一种融洽和谐的关系及相互的信任感，这些都是回应者回答敏感性问题所需的。

以上的说法都有道理。自填问卷法与访谈法相比，前者所得的关于敏感话题的资料往往更为准确。与有访谈员参与的方式相比，自填问卷法无论是使用笔答还是电脑辅助，都会产生更准确的结论（Acquilino，1994；Dillman & Tarnai，1991；Tourangeau & Smith，1998；Turner et al.，1998）。并且，对十分敏感的内容（例如，毒品使用和性行为）以及涉及个人形象的更微妙的问题（诸如健康问题或在前列腺手术后的“普遍问题”）的调查中，也显示出同样的效果（Fowler，Roman，& Di，1998；McHorney，Kosinki，& Ware，1994）。如果敏感问题是调查关注的焦点，那么运用一种没有访谈员参与的方法来获取答案肯定会提高判断的精确度。

从这个方面来比较电话和面对面的访谈时，其区别就不那么明显（De Leeuw & Van der Zower，1988）。然而，当问题的答案暗含一种社会偏见倾向时，使用电话访谈通常就会与面对面访谈表现出多方面不同。曼吉温（Mangione，1982）等人发现，当调查人们过去有酗酒问题的比例时，用这两种方式得到的结果有明显差异。霍奇斯汀（Hochstim ，1967），亨森、罗斯、坎内尔（Henson，Roth，& Cannell，1977），艾奎利诺（Acquilino，1994），福勒、罗曼和迪（Fowler，Roman，& Di，1998）也发现了相同的结论。

问题内容影响收集资料的方式的另一个全然不同的方面是问题回答的难度。在一些调查中，研究者可能会问到一些很难准确回答的事件或行为，原因是时间过去了很久或是事情太琐碎。这种情况下，报告的准确度可以通过查询记录或与家庭中的其他成员进行讨论而得到提高。标准化访谈是一个快速的问答过程，它很少提供这些机会，尤其是在电话访谈中。自填问卷法为调查对

象提供了更多的考虑、查找记录以及与家庭成员商讨的时间。

当需要了解关于某些小事件的细节性问题时，例如人们吃什么东西、怎样花钱、看了什么电视节目等，最好的办法就是叫回应者记日记。虽然这要受到记日记人数的限制，而且研究者也会给其一定的报酬，但是这不失为获得细节的好方法。有一种代替书面日记的方法就是使用电脑辅助，其中包括让人们拨打 800 电话或在网上提交，同样能获得高质量的数据。

总地来说，当样本具有可比性时，研究者发现很多的调查结果不会受到资料收集方式的影响。除非刚才提到的那些问题在研究项目中特别突出，收集资料的方式应根据研究的课题来确定，而不是以问题单与资料收集方法间的相互影响为转移。不过，注意问题的形式或内容与资料收集方式间的相互作用对调查结果的影响，仍然是调查设计过程的一个重要组成部分。

回应率

在资料收集方式的选择中，对回应率的考虑比其他因素更为重要。显然，当使用团体填答调查法时其巨大优越性之一就是高回收率。一般说来，要求在教室的学生或在工作环境中的工作人员填写问卷时，回收率几乎是 100%。对回收率的影响通常是由旷工旷课或按时刻表轮班工作或休假造成的。

邮件或电子邮件调查的回应率主要取决于研究的总体和调查的目的。在由美国老年保健医疗机构进行的关于前列腺手术病人的调查中，其邮件回应率高达 82%。在给无回应者打电话后，其回应率超过了 90%(Fowler，Roman，& Di，1998)。当然，这类邮件调查中也有回应率少于 20%的情况。

毫无疑问，无回应的情况是与使用邮件调查的方法有极大关系的。在前一章中提到过，如果研究者只简单地给选取的调查对象寄出问卷却没有再寄一封信或发一份问卷，其回应率大多小于 50%(Heberlein & Baumgartner，1978)。如果采用了广泛而合理的补救措施，并且对调查进行了精心设计及认真执行，邮件调查的回应率就会与其他方式获得的回应率相似(例如，Dillman，2000)。

电话调查的高回应率部分地取决于抽样的方式。使用电话进

行调查的一个方法是复现个人访谈。如果研究者除了有电话号码,还有通讯地址,就可以先寄一封信,用来介绍研究项目和解释研究目的。然后访谈员可以给调查对象打电话以请求合作。在这种情况下,电话和个人访谈的回应率不会有显著的差异。尤其是如果访谈员亲自登门造访无回应者时,情况更是如此(Groves,1989;Hochstim,1967;Mangione 等,1982)。

然而这里提到的方法只代表了一小部分的电话调查。更多的电话调查是与随机电话拨号抽样相联系的。随机拨号的一个独特之处在于:因为不知道地址,不可能预先通知回应者。即使是有能力并尽责的调查机构所进行的电话调查,其回应率也比通过个人访谈获得的回应率低 5% 以上(例如:Groves & Kahn,1979;Groves & Lyberg,1988)。

在一些城市地区,因为使用了电话而使回应率有所提高。然而在郊区和农村地区使用电话获得的回应率比使用个人访谈员的回应率低。使用电话可以更好地增强对住在有安全系统的楼房里的家庭的调查,因为访谈员晚上不方便去那里。在电话调查中,也可以打 10 个甚至是 15 个回访电话去找那个极少在家的回应者。从另一方面来说,正如在第 3 章里提到的,在随机电话拨号调查中有些无回应的情况是与其教育状况和年龄有关的,这种情况在个人访谈调查中则出现得相对较少。

总之,出于对回应者的尊重,如果先寄信告知情况,电话调查方法和个人访谈的方法之间的区别就很小。另外,当研究者想对先前访谈过的人进行回访以进一步获得资料时,利用电话与通过个人访谈的回应率没有什么差别。在一些人口稠密的市中心区,电话调查方式,即使是采用随机电话拨号,获得的回应率与个人访谈几乎一样(甚至会更好)。

就更大范围的人群而言,由于随机电话拨号调查回应率比通过个人访谈调查回应率低,这是导致使用随机电话拨号调查成本高的原因之一。当回应率下降 5% 到 20% 时,就要乘以排除在样本之外的没有电话的人的比例。回应率的差别不是琐碎而无价值的,当研究者选择随机电话拨号方法时,必须有准备接受其缺陷,或者努力避免其缺陷。

成 本

邮寄和电话调查的诱人之处是它的单位成本比个人访谈的低。当然,成本最低的调查是使用互联网。调查成本取决于多种因素,一些较突出的因素是设计问卷所需的专门时间、电脑辅助程序的编制和调试时间、问卷长度、样本的地理分布状况、样本的可用性和利益、回访的措施、选择回应者的原则以及训练有素的工作人员的可得性。

人们可能会误认为邮件调查费用低,其实,邮资、工作人员邮寄问卷时间的成本以及印刷问卷的成本是一个不小的数目。另外,如果要进行电话提醒,成本将更大。

成本对比中另一个关键是有关电话费用。使用电话的成本也会影响到个人访谈调查与电话调查二者成本的比较。但是,在同一样本的调查中,多数情况下,个人家庭访谈调查比电话调查的成本高。访谈员走访家庭以及联系调查人的工资和花费都必然会远远超过电话调查。

邮件调查和电话调查的成本对比也取决于总体状况。对于一个积极性十分高、能迅速地对调查作出反馈的样本来说,邮件调查成本就会低于电话调查的样本。然而,如果只想获得相似的回应率,运用邮件调查或电话调查的费用基本相同。

尽管在选择采用邮件调查还是电话调查时可以不考虑成本,但在确定是否选择个人访谈调查方式时,成本却起着重要的作用。不过,在许多案例中,个人访谈的优点使其成为达到一系列研究目标的最好方法。

最后,如果调查可以在互联网上进行,那么其成本将是最低的。根据收集资料的类型,可能需要对软件进行一些先行投资。除此以外,则是设计和测试调查工具所需工作人员的时间成本。当然,与其他调查方式相比,问题的关键在于是否可以通过电子邮件联系到合适的样本以及所获得回应率的大小。

工具的易得性

在选择资料收集方法时必须要考虑工具和工作人员的易得

性。培训一个访谈员费时费力，新访谈员培训耗损率一般较高。很多新的访谈员不善于赢得回应者的帮助，刚开始时被拒绝率会很高。另外，善于培训和指导访谈员的人也不容易找到。因此，对于那些想进行有访谈员参加的调查来说，最实际的考虑就是访谈员收集资料的专业能力。如果有人曾经进行过贯穿整个过程的调查或者有指导和培训访谈员的经验，那么利用访谈员的调查就较可行。否则，选用自填问卷法将更有利。

资料收集的时间跨度

资料收集的时间是因方法而异。邮件调查通常要 2 个月才能完成。其常规程序包括邮寄问卷、等待、再多寄一些问卷、再等待，最后是打电话或派人上门进行进一步的访谈。当然，网上调查不考虑等待邮件邮寄的时间，但是它通常需要反复进行联系和提醒。而电话调查可以在几日内完成。最快的调查方法要为无回应付出代价，因为有些人很可能在短期内找不到。但是电话调查通常可以比相应范围的邮件或个人家庭访谈调查快。

做个人家庭访谈调查的时间长度不能一概而论，因为它决定于样本范围和工作人员的易得性。但是可以肯定的说，只是在十分特殊的情况下个人访谈的资料收集时间短于电话调查所需要的时间。

电脑辅助的资料收集

传统的调查是依靠访谈员或回应者手中的纸和笔进行的。然而，在过去的 10 年里，电脑取代了纸和笔：调查的问题都被输入电脑，访谈员或回应者通过屏幕阅读，答案通过键盘输入电脑。电脑辅助资料收集的主要优点在于，答案当即就被储存为机读形式。相对其他调查而言，这种方法有许多优点：

- 电脑可以按照复杂问题跳越的方式进行，这是纸笔形式很难办到的。
- 从前面的问题甚至从前面的访谈中得到的资料，可以在问题的表述或者问题提问的次序中加以考虑。
- 如果被提供的是不连续的资料，那么电脑能识别出不连续

性,并使之符合资料的收集。

其缺点在于,需要相当长的时间来确保电脑辅助资料收集没有错误,并且如第8章将更为详细地谈到的那样,研究者没有能力对资料输入过程的质量进行检查和控制。因此,就像大多数关于调查设计的决策一样,电脑辅助资料收集的价值是因不同项目的要求而异的。萨瑞斯(Saris,1991)详细地描述了电脑辅助访谈的成本和优势。库珀(Couper,1998)等展示了大量的与电脑辅助资料收集有关的最新研究。

在进行电话调查时,以电脑予以辅助是十分普遍的。事实上,对一些人而言,电脑辅助电话调查(CATI)从本质上讲是与电话调查同义的,尽管有很多电话调查仍然使用纸和笔。迄今为止,除了在减少资料丢失方面外,没有任何文献认为电话调查得来的资料的质量会因使用电脑辅助而改变(Catlin & Ingram,1988)。大多数研究指出的优点和缺点都是很切合实际的:便于问题形式和顺序的处理、资料输入的速度、样本管理以及以指令或所需定义的形式为访谈员提供"帮助"的可能性;缺点是它需要一个完全无误的程序、访谈员修正某些问题的困难以及存在电脑系统崩溃的风险;另外,虽然叙述式问题可以通过由访谈员逐字打出答案的方式予以处理,但是电脑辅助资料收集法进一步增加了提问只有固定回应项的压力。

大多数电脑辅助访谈都是通过一个集中电话设备进行的。然而,随着轻便的手提电脑的出现,访谈员进行个人家庭访谈时运用电脑辅助资料收集成为可能。另外,在诸如医生办公室等选定的环境里,人们可以通过电脑自填问卷的方式收集资料:回应者坐在电脑前,阅读屏幕上的问题,然后在没有访谈员的情况下自行输入答案。那种有触摸式屏幕或者有鼠标功能的电脑特别适用于这样的资料收集方式。电脑辅助个人访谈(CAPI)的优点与电脑辅助电话访谈(CATI)几乎一样:便于问题的处理和资料的快速编辑。一个访谈日结束后,访谈员可以通过电话线传送资料以便集中保存处理。

尽管电脑辅助自填问卷调查的技术还在发展中,但其中还有更多令人感兴趣的潜力将被人们认识到。例如,通过电脑我们可

以使用有别于文字的形式来呈现信息和刺激物（例如，图画）。电脑不但可以为那些有一定阅读障碍的人读出问题，还可以调整问题的语言以适应回应者的语言。电脑具有改变问题的选择或次序以适应先前应答的能力，这是在具有复杂指令的自填问卷调查中运用电脑的独特优势。交互式电脑可以提问并通过按钮式电话资料输入机记录答案，这就为网上调查提供了一种新的选择方式，使回应者能在他们觉的方便的任何时候提供资料。最后，回应者在回答敏感性问题时会觉得输入电脑比面对访谈员更自在。

显然，成本更低、方便携带以及功率极大，电脑的这些发展为研究者提供了一种收集资料的新资源。尽管将电脑用于电话总机式访谈是最简单的使用方式（这是当前最先进的），但是，电脑运用于家庭访谈和自填问卷调查（无论亲自或是通过网络）将是十年内发展最迅速的领域。就像任何调查设计的特点一样，电脑辅助能解决许多问题，同时它又产生了一些新问题。哪种设计是最好的最终取决于特定调查的需要。

各种方法的简要比较

前面的讨论并非详尽无遗，但是它涵盖了大多数的主要因素。资料收集方式的选择是非常复杂的，它牵涉到调查研究过程的很多方面。接下来将总结各种资料收集方法的一些优缺点。

个人访谈的优点

- 有些样本只有通过个人访谈才能得到最好的效果（例如，区域概率样本）。
- 对于大多数总体来说，个人访谈可能是取得合作支持的最有效途径。
- 访谈员处理的优势，诸如回答回应者的问题、探寻充分的回答、准确执行复杂的指令和顺序等。
- 可以使用多种方式结合进行资料收集，包括使用笔录和电脑的观察、视觉提示和自我填答。

- 可以建立和谐和信任的气氛(包括为获得某些非常敏感的资料而写的书面保证)。
- 比起其他方式,可以使用更多的调查设备。

个人访谈的缺点

- 与其他方式相比,成本相对较高。
- 需要在样本附近地区训练有素的访谈人员。
- 资料收集的总时间可能比电话访谈的方式长。
- 有些样本可能不如通过其他方式更易接近(如那些住在超高层建筑或高犯罪率地区的人、社会名流、职员、学生等)。

电话访谈的优点

- 与个人访谈相比,成本更低。
- 可以对一般总体进行随机电话拨号抽样。
- 与个人访谈相比,更易于接近特定的群体。
- 资料收集时间较短。
- 有利于访谈员对访谈过程进行控制(与邮件调查或互联网调查相比)。
- 与个人访谈相比,对访谈员的雇佣和管理更容易:所需工作人员更少,不一定要在样本附近,能更好地进行指导和质量控制。
- 与邮件资料收集方式相比,回应率更高。

电话调查的缺点

- 抽样具有局限性,特别是排除了那些没有电话或者因为没有获得正确的电话号码而无法联系到的人。
- 与个人访谈相比,RDD 抽样产生的无回应率要高。
- 存在问题编制或测量的约束,包括在回应的可选择性、可视辅助的使用和访谈员的观察等方面的局限。
- 可能不大适合涉及隐私或敏感性问题的调查。

自填问卷法的优点(与有访谈员参与的访谈相比较)

- 易于利用可视辅助来呈现问题(与电话访谈相比)。
- 所提问题可以有篇幅长或复杂的回应项类型。
- 可以问一系列相类似的问题。
- 回应者可以不受访谈员影响而独立回答问题。

自填问卷法的缺点

- 需要特别仔细的问卷设计。
- 对开放型问题不是很有效。
- 回应者需要有较好的阅读和书写能力。
- 没有访谈员对问题回答过程的全面性、合目的性的质量控制,或者说答案质量没有控制。
- 不能对由谁来回答问题进行控制。

自填问卷可以通过邮寄问卷、团体调查或者在家庭中进行调查。每一种方式都有其优点和缺点。

团体调查的优点

- 通常具有高的合作率。
- 有机会解释所研究的项目和回答与问卷有关的问题(与邮件调查相比)。
- 成本很低。

主要的缺点是:只有少数的调查能够使用团体样本。

邮件调查法的优点

- 相对来说成本较低。
- 可以在极有限的工作人员和设备的情况下完成。
- 可以取得分散的和很难通过电话调查或面谈方式联系到的样本。
- 回应者有时间仔细思考答案、查阅记录、询问他人。

邮件调查法的缺点

- 邮件作为获取合作方式的无效性(取决于所研究的组群和主题)。
- 在资料收集中存在没有访谈员参与的各种缺点。
- 需要样本的详细地址。

给各个家庭散发(过后回收)问卷的优点

- 与邮件调查法相比,访谈员可以对研究项目作出解释、回答问题和选定家庭中的一员作为回应者。
- 与个人访谈调查的回应率几乎相同。
- 与个人访谈和电话访谈相比,回应者能更仔细地思考答案并且有更多的机会查询记录或询问家人。
- 不需要训练有素的访谈人员。

给各个家庭散发问卷的缺点

- 这种方法的成本与个人访谈的相当。
- 需要现场工作人员(尽管不要求像个人访谈那样经过精心训练)。

互联网调查的优点

- 收集资料的单位成本低。
- 回收速度快。
- 具有自填问卷法的所有优点。
- 具有电脑辅助法的所有优点。
- 像邮件调查一样,为回应者思考答案、查找记录或询问他人提供了足够的时间。

互联网调查的缺点

- 样本局限于互联网的使用者。
- 需要详细的地址。
- 面临获得合作的挑战(取决于被抽样到的群组和研究主题)。

• 具有没有访谈员参与的资料收集方法的各种缺点。

最后,在选择方案时,研究者也应该考虑各种资料收集方式的结合。如前所述,很多问题的答案不会受资料收集方式的影响。把个人、电话、邮寄和互联网资料收集方法结合在一起,可能会节约总的成本,因为有的资料收集方式不需要有关抽样或无回应率的花费。迪尔曼(Dillman,2000)对此做了一些讨论。

总　结

显然,资料收集方法的选择是一个复杂的决策,它在很大程度上取决于特殊的研究情况。所有提到的策略对某些研究而言是最好的,然而需要注意的是,在对普通的以家庭为基础的样本的调查中,情况会有所不同。25 年前,有研究者可能会认为个人访谈调查是大多数研究选择的方法,但是,这种观点难以说服那种认为别的方法也能获得令人满意资料的争辩者。

研究者必须直接说明不能用电话访谈取代访谈员访谈的原因在于节约成本。尤其出于对主题、测量内容和无回应问题的考虑,研究者认为有些研究更宜于采用个人访谈法。然而,公正地说,这时研究者在解释不用电话调查的理由,而没有解释采用电话调查的理由。

自我填答技术在近十年里获得发展的原因有二:首先,是互联网的发展为资料收集开辟了一种新方法。研究者期待对互联网潜能的进一步开发。其次,是相当多的研究证明,自填问卷方法,尤其是有电脑辅助的方法,与有访谈员参与的方法相比,可以更好地收集到有关敏感话题的资料。这些发现伴随着研究范围的拓展,如毒品使用和危险的性行为,使人们产生了整合这些资料收集方法的兴趣。

最后,应当明确,在确定资料收集方法时,整体的调查设计方法是很重要的。与花费了同样的代价的较大规模的电话访谈样本相比,较小规模的个人访谈样本有可能提供更多的有效资料。在决定如何收集调查资料之前,具有一个正确的方法目标观,并对影

响成本和资料质量的所有设计方面的问题进行缜密考虑，显得十分重要。

练　习

不考虑金钱成本，设计一个使用邮件方式为最佳选择的调查，并解释使用邮件比其他方法好的原因。然后以同样的方式对随机电话调查、个人访谈家庭调查和互联网调查方法进行设计。



5

良好测量问题的设计

Designing Questions to Be Good Measures

在调查中，对答案的兴趣不仅仅是因为答案本身，而是因为它们与被测量对象有联系。好的问题具有信度(在类似的环境中提供一致的测量数据)和效度(答案符合问题所要测量的对象)。本章将从理论和实践方面探讨设计具有信度和效度的问题的方法。

为某种调查手段设计问题实际是在设计一种尺度，而不是在进行一种对话式的调查。一般说来，调查所得的答案并非其本身使人感兴趣，答案的价值在于它能显示出一个可预计的令人感兴趣的客观事实之间或主观状态之间的关系的程度。好的问题能够使其答案与研究者试图测量的内容之间达到最大相关。

从某种意义上说，答案只是由研究者精心设计的人为环境所引发的回应。本章要讨论的关键问题是，针对我们所关注的事实，一个调查问题的答案能告诉我们一些什么。现在我们来看一下一些答案和它们所代表的意义：

- 一个回应者告诉我们，在 1996 年的总统选举中他投了多尔的票而不是克林顿的。我们感兴趣的是在投票站他拉动了哪一个控制杆。由于种种原因，调查的答案可能会与投票站中实际发生的事有很大的差异。回应者可能会拉动错误的控制杆，因而不知道他选举的到底是谁；回应者也有可能忘记了他投了谁的票；回应者也许由于某种原因而故意改变自己的答案。
- 一个回应者告诉我们在过去一年里他去医院接受治疗的次数。这与研究者在过去一年里每天 24 小时跟踪得来的数字是一样的吗？对于回忆产生的问题、对看医生的定义的问题、准确报告的意愿问题，都会影响到回应者所给出的与研究者独立调查所得出的次数之间的不一致。
- 当一位回应者认为公立学校制度“好”而不是“一般”或者“差”时，研究者希望能通过这个回答反映出对那个学校制度的评价和理解。假如回应者评论的是一所学校（而不是整个学校制度）、改变了答案以迎合访谈员或者对该问题的理解与他人不同，那么，其答案可能没有反映出研究者试图测量的感受。

大量调查研究被人们分析阐述以说明研究者能了解回答的确切意思。针对评估回应者答案和客观事实之间一致性的研究表明，许多回应者能很好地回答大多数问题。然而，认为人们给出的答案和事实十分吻合的想法却是非常天真的。要使答案成为很好的测量尺度，通常需要精心的设计。接下来将要探讨研究者用来

提高回应者答案和客观事实的一致性的方法。

良好测量的一个目标是提高问题的信度。在相同的条件下，两个回应者会以同样的方式来回答问题：回应者之间存在不一致性的程度、随机误差的引入以及测量工具的不精确性。本章的第一部分将讨论怎样增加问题的信度，还要讨论给出的答案与研究者所试图测量的目的之间的关系问题：答案的一致性有多高？本章的最后两部分会谈到效度问题，包括答案和客观事实间的一致性问题和提高一致性的方法问题(Cronbach & Meehl，1955)。

提高答案的信度

确保测量一致性的一个措施是使样本中的每个人回答相同的问题，并将答案记录下来。研究者会期望做出这样的假设：答案差异来自回应者的差异，而不是来自每个回应者所受到的那些刺激的差异。显然，问题的措辞却是那些刺激的中心部分。

- 调查资料的收集是研究者和回应者之间的一个互动过程。在自填问卷的调查中，研究者通过书面问卷或电脑屏幕与回应者进行直接交流；在另一些调查中，访谈员为回应者读出问题。在每个调查中，调查手段都是为调查对象提供问题。为了使所有的回应者具有相同的资料收集经历，好的问题应具有以下特点：
- 研究者的提问是完全编写好了的，以便这些书面的提问为回应者回答问题做好了充分的准备。
- 对于每个回应者来说提问的方式都是相同的。
- 构成对问题恰当回应的各种答案应一致地传达给所有的回应者。

不恰当的表述

关于问题中存在不恰当表述的最简单的例子，就是研究者的表述不能构成一个完整的问题。

表述不完整的例子

	不好的	好的
例 5.1	多大?	在你过上个生日时你的年龄是多大?

访谈员(或回应者)必须自己加上一些词或者改变词语,以使上面左边的问题便于回答。如果目标是让所有的回应者都回答同样的问题,那么研究者最好是将问题完整写出。

有时要选择恰当的表述以适应不同回应者的具体情况,但这并不意味着研究者不必写出问题。一般是把可选择的表述放在括号里,当这些表述符合实际情况时,访谈员就使用它们,不需要时就省略去。

选择性表述的例子

例 5.2　在过去一年里,你(或住在你周围的人)曾经被陌生人攻击或殴打吗?

例 5.3　(你/他/她)向警察报过案了吗?

例 5.4　在(你/他/她)的上一个生日(每一个人)的年纪是多大?

在例 5.2 中,如果研究者已经知道回应者是独居的,那么括号里的插入语就会被省略。否则,研究者将保留它。在例 5.3 里,括号内的选择看起来是不重要的,但是,括号里的词语提醒访谈员必须对表述作出选择;要使用正确的代词;为了呈现一个令人满意的刺激,访谈员应完全依据所给出的内容念出问题。

在例 5.4 里,以不同的形式完成了同样的任务。如果以同样的问题用来向家庭里的每个成员提问,这种格式就会被使用到。与其不断地重复同样的话,不如只写一个问题,并且指导访谈员用合适的词进行替换(你的丈夫/你的儿子/你的长女)。

当然,电脑辅助手段的一个优势是可以用电脑填入合适的词语,而不是由访谈员根据各种情况调整问题的表述。不管是写在纸上还是通过电脑,目标都是让访谈员提出清楚的问题,并利用访谈员在以前的访谈中所获得的知识,根据回应者的不同情况对问题作出修改。有一种偶尔出现在问卷中的不正确的选择性问题表述:

不正确的选择性表述的例子

例 5.5　在你住所附近你最喜欢什么？（我们对任何事情都感兴趣，比如房子、人、公园或者所有的一切）

或许使用括号插入的形式对回答问题有困难的回应者来说是有所帮助的，但是从测量角度来看，它破坏了标准化访谈的原则。如果回应者不能立刻想出答案时访谈员使用了括号插入的方式，就相当于不同回应者回答了不同的问题。这种选择性的插入表述是在研究者认为最初的问题不是十分完善时引入的。合理的方法是把编写一个完善的问题放在首位，除非为了访谈员按照标准化方式处理问题以适应回应者的具体情况，否则，不应给访谈员选择读哪些问题或怎样读它们的权利。

下面是关于不准确表述的另一个例子，在这个例子中有三个错误的地方：

措辞不准确的例子

例 5.6　我希望你能通过使用非常好、好、一般或差这四个等级来评估你周围事物的不同特点。在我念出以下选项时请认真思考。

A. 公立学校　　B. 公园(parks)

C. 公共交通　　D. 其他

例 5.6 中存在的第一个问题是题干的顺序问题：选择项的阅读先于关于仔细思考特定项目的引导语。回应者可能会忘记问题，因此访谈员可能不得不在回应者做出回答之前做一些解释或重复表述。存在的第二个问题是，调查员对第二个选项中的"parks"一词没有加以限定（因为，在英语中，parks 有多种译法，如公园、停车场、运动场等，容易产生歧义——译者注）。更好的表述应该是：

表述准确的例子

例 5.6a 我想问你对周围事物不同特点的评价。我希望你能仔细思考你的答案。你是怎么评价的——很好、好、一般或者差?

这就考虑到了访谈者首先想了解的问题,以及所有的选项。

例子 5.6 中存在的第三个问题是,对于第 4 选项"其他",访谈员会说什么?可以临时编造一些新的问题吗?例如"你还有对周围事物的其他评价吗?"这个不确定的问题如何表述?类似例子中这种"其他"选项在问题单中出现的情况还比较普遍。显然,例 5.6 这种形式的问题表述是不恰当的。

上述例子说明:如果表述不完善,就不能保证对所有回应者所做的提问是一致的。增加一致性的另一个步骤是编制一系列流畅而又简明的问题。如果问题使用的表述令人费解或含混不清,有些词语很难读或者连成句时感觉很累赘,访谈员就要换一些词以使问题听上去更顺耳或者更方便读出。虽然可能通过对访谈员进行训练和指导也能达到这种要求,然而,最明智的办法还是尽可能给访谈员容易读的问题以便于他们提问。

确保回应者对问题的理解相同

如果对所有的回应者问完全一样的问题,已经采取的步骤是确保答案的差异归于回应者的差异。但这里还需考虑的事情是:对所有的回应者而言,问题的意思都应该是一样的。如果两个回应者对问题的理解不同,那么,他们的答案就可能因此而不同。

产生问题的原因是大量使用了难以理解的词语。因而,对样本受教育的程度和文化背景的关注将会集中体现出来。即使面对的是受过良好教育的回应者,也要使用那些短的和容易理解的简单词语,这在问卷设计中是非常重要的方法。

毫无疑问,比使用不熟悉的词语更容易犯的错误就是使用专业术语或者容易产生歧义的概念。有研究对一些在日常生活中经常使人产生误解的词语做了详细的论述(例如,Belson,1981;Fowler,1992;Oksenberg, Cannell, & Kalton,1991;Tanur,1991)。

界定不明的术语

例 5.7　去年，你看医生或者与医生讨论你的健康状况的次数是多少？

这个例子中存在的问题是，表述中有两个模糊的术语或概念：第一是医生的概念不明确。是不是只包括那些从事医药实践并具有医学博士学位的人？假如这样的话，精神病学家被包括在内，但是心理学者、脊椎指压治疗者、整骨疗法家和足病医生则被排除在外。医生的助手或者在医生办公室直接为医生工作的护士算不算？如果一个人去医生的诊室接种疫苗，但是是由护士执行的，这种算不算在内？第二，去见医生或者与医生讲话的内容是什么？打电话咨询算不算在内？去见医生但是医生不在，这算不算？

解决这个例子中存在问题的最好办法，通常是向回应者和访谈员提供他们所欠缺的定义：

例 5.7a　我们将要询问的是你见医生和从医生那里得到的医疗建议。在这种情况下，我们感兴趣的医生是所有的具有医学博士学位的人或者在办公室直接为他们工作的专业人员，例如护士或医疗助手。

正如在这个问题中所遇到的情况，当定义极其复杂并费时较多时，可采用另一种方法：对很可能被遗漏的事情提供补充性的提问。例如，对见精神病学家、接种疫苗以及电话咨询经常报告不够，这就需要确保有专门的后续问题作为补充。

界定不明的术语

例 5.8　昨天你吃早餐了吗？

该例的问题在于，早餐的定义相当宽泛定义困难：有的人认为在中午以前的任何时候喝的咖啡和吃的油炸圈都算是早餐；另外一些人认为那不算吃过早餐，除非包括火腿和鸡蛋，并且要在早上八点以前吃完。如果目的是测量食品的消费量，那么结果就可能会由于对早餐的定义不同而出现巨大误差。

解决方法有二：其一，是选择对早餐的概念进行定义。这通常是一个好方法，但是却十分复杂。其二，与向回应者传达一个普遍的概念相反，研究者可以只要回应者报告他们昨天在 10 点以前吃了些什么。这不要求每个回应者对早餐的定义是相同，只要在编码阶段把他们吃的食品归纳到一起看是否符合早餐的标准即可。

例 5.8a 基于研究的目的，我们把早餐当作一顿正餐(meal)，在早上 10 点以前吃；包括一些蛋白质的食品，例如鸡蛋、肉类或者牛奶；包括一些谷类食品，例如烤面包或者谷类早餐食品；以及一些水果或者蔬菜。根据这些定义，请问你昨天吃早餐了吗？

界定不明的术语

例 5.9 你是赞成还是反对枪支管制立法？

例 5.9 中存在的问题：枪支管制立法可能指禁止某些种类的枪支的合法销售，让人们给所持枪支注册，限制人们所拥有的枪支数量或种类，或者限制拥有枪支的人。如果不假定回应者对这个问题的理解，答案就无法解释。毫无疑问，回应者将会以不同方式阐释这个问题。

例 5.9a 有一个控制枪支的议案是这样的，不允许任何被证明有过暴力犯罪记录的人购买或拥有手枪、步枪或者鸟枪。你是反对还是赞同这样的立法呢？

有人可能会争辩说这只是枪支管制多种议案的一种。确实是这样。如果有人要问有关枪支管制的各种方法的复合问题，那么，他就应当把问题分解成多个明确的问题，以便能被所有回应者理解、能被研究者解释。研究者不解决好这样的复杂问题而留给回应者自己来决定他们要回答的是什么问题，这样做是无济于事的。

在提供给所有回应者一个复杂的概念与保持问题的简单明了之间存在一种潜在的张力。这种情况在访谈员参与的调查中特别

突出，因为以口述的方式传达时，对一个很长的定义是难以把握的。

有一种可能的方法，是告知访谈员，在回应者要求明确问题或对问题产生误解时才提供定义。值得注意的是，这种方法中如果定义是访谈员临时提供的，那么他们提供的定义能否前后一致。不过，电脑辅助访谈使得为访谈员提供精确的定义变得容易起来。用这种方法时，更需注意的另一个问题是，只有一部分回应者才得到了所需的定义，而那些没有要求明确问题或对问题没有显示出误解的回应者，会缺少那些很可能影响他们答案的重要信息。

在康拉德和斯科伯（Conrad and Schober，2000）的实验中，给访谈员根据回应者的需要提供定义和解释问题的自由。虽然有事实表明精确度确实提高了，可是随之而来的是访谈员培训费用的增加和访谈时间的加长。针对更多的研究需要解决如何就复杂概念进行提问的状况，适用于大多数调查的最好方法是：避免使用复杂、模糊的术语，并对那些在问题表述中用到的术语进行定义。

避免复合问题的出现

另一种导致问题不具有信度的情况是一题两问。

例 5.10　*你希望自己富有并出名吗？*

问题很明显："富有"和"出名"是不同的，可能有的人只希望拥有其中的一个方面，当面对这样的两个问题时，回应者不得不自己决定回答哪一个，而且取舍也因回应者的不同而不同。

大多数的复合问题有些难以捉摸，如：

例 5.11　*在过去的 30 天里，当你从 ATM 柜员机上取款时，你取款少于 25 美元的频率——总是、通常、有时、从不？*

这个问题要求进行 3 次以实际经验为根据的分析：计算去

ATM 上取款的次数，计算取款少于 25 美元的次数，以及两者之间的关系。尽管从严格意义上来讲这只是一个问题，但是它需要预先回答两个问题才能得出答案。如果将其设计成两个问题将会好点。

例 5.11a　在过去的 30 天里，你从 ATM 柜员机上取款多少次？
例 5.11b　（如果取过）取款少于 25 美元的有多少次？

例 5.11a 和例 5.11b 问题组的另两个优点值得注意：第一，它区分了那些从不使用 ATM 柜员机的人，因为这个问题不适合那些人。第二，两个问题都使用具体数字，避免了让回应者做计算的麻烦。简化对回应者的要求，几乎总是一种好的提问方法。

例 5.12　你到哪里进行日常医疗保健？

这个问题是假设所有的回应者都进行了日常医疗保健，但这个假设是不准确的。它应该以两个问题的形式提问。最好的提问形式或许是：先问在某一时段是否进行了日常医疗保健（例如，过去的 12 个月里），如果回答是肯定的，再问在哪里进行保健的。

“不知道”选项

当回应者被问到有关他们自己的生活、感受或经历时，“不知道”选项通常用来表明他们不愿意作出的回答，或者所提的问题是他们确实不知道的。如果问题的主题远离他们当前的生活，很多回应者将会没有足够的知识以形成一个见解或想法。在这种情况下，常有一题两问的例子：你有回答这个问题所需的信息吗？如果有，那么你的答案是什么？

有两种方法可以处理这种可能的情况：其一，直接向所有回应者提问，让回应者自愿说出“不知道”的答案。但是，回应者在自愿回答“不知道”时所隐含的意思是不一样的（Schuman & Presser，1981），访谈员处理“不知道”选项时的方式也不一致（Fowler & Mangione，1990；Groves，1989）。其二，问所有的回应者一个标准

化的筛选性问题，即问他们是否对某个主题足够熟悉以至能对它有一定的见解或想法。

当回应者对主题很熟悉时，是否要进行筛选性提问并不重要。当有相当数量的回应者不熟悉或没有想过那个主题时，不管提问的是什么问题，最好要先对相关主题做一个熟悉程度的筛选性提问。

为特殊的子群设计的特殊表述

研究者已经致力于解决不同子群具有不同词汇的问题。有人可能认为问题在于标准化的测量需要对不同的子群使用不同的问题(Schaeffer，1992)。

但是，为不同的子群设计不同的问卷形式似乎从未做到过。方法论学家致力于找到适用于整个总体的具有一致含义的表述。即使一个问题的表述对一部分人(大多具有良好的教育)更具有语言上的典型性，但是，要找到准确可比的表述方式以便适合总体中的另一部分人，并随后给访谈员可靠的标准以便让其决定什么时候用哪一个版本的表述提问，是十分烦琐困难的。可能其导致的不可靠性之大，还不如干脆不考虑这个问题。

最大的挑战是怎样从说不同语言的人群里收集具有可比性的资料。最谨慎的做法是，找一位翻译把原文翻译成另一种语言的新版本，然后让另一位翻译把新版本回译成原来的语言，再尝试消除原文和回译后得到的版本之间的不同。

如果问题的设计者最初就考虑到要便于翻译，那么这个过程就能被极大地改进。例如，对数字的翻译就比对形容词容易；想要把抽象的概念和口头语翻译准确就十分困难。即使十分谨慎地进行翻译，也很难保证人们用不同语言回答的问题具有可比性。不同的语言中，形容词所表达的规模的可比性是令人怀疑的。问题越具体，不同的语言和文化间的可比性就越高。马林 G. 和马林 B. V. (Marin & Marin，1991)很好地分析了从讲英语和讲西班牙语的人那里收集可比资料所面临的挑战。

回应类型的标准化期望值

如上所述，给访谈员一个好的文稿以便他们能准确按照表述读出问题，以及所设计的问题能让所有的回应者理解一致，这些都是重要的。但是，好的问题还有一个关键组成部分，那就是回应者应该对问题的合理答案的构成有一个共识。

让回应者都有这样一个共识的最简单的办法，是提供给他们一些可选择的答案。这样的问题叫作封闭型问题。回应者从研究者提供的答案中选出一个或多个答案。

封闭型问题并不适合所有的情况。与提供的答案相比，问题本身的答案可能范围更加广泛，但这些答案并不是都可以预测。因此，研究者会倾向于不给回应者提供可选择的答案。这样，所提的问题就必须尽可能地传达出期望得到的回答种类。

例 5.13　*你什么时候得的麻疹？*

例 5.13 中存在的不足是，这个问题没有详细指明回应者应回答的术语。可能会有的答案："5 年前"、"当我在部队的时候"、"当我怀上第一个孩子的时候"、"当我 32 岁的时候"、"在 1987 年"。所有这些答案都可能是同一个人给出的，而且所有的答案都可以被看作是对问题作出了回答。但是在同样的调查里，这些答案并非都是可以接受的，因为描述性统计资料要求具有可比性的答案。访谈员使用例 5.13 里边的措辞不能得到一致的可比性资料，因为每个回应者都必须猜测所需答案的类型。

解决上述问题的办法是，编制新的问题以便给回应者解释所需答案的类型：

例 5.13a　*你在多大的时候得过麻疹？*

很明显，例 5.13a 的提问方式才是研究者应为所有回应者提供的问题表述方式。

例 5.14　你为什么投候选人 A 的票？

例 5.14 中存在的问题是几乎所有的“为什么”的提问都会引起的一类问题。原因是一个人的随意性或参照标准会影响答案。在这个特定的例子里，回应者可能会选择谈论候选人 A 的优势、候选人 B 的弱势，或者谈其选择某种标准的理由（如，我的妈妈一生都是共和党人）。由此，即使观点一致的回应者也可能作出不同的回答。

解决这个问题的办法是明确问题关注的重点：

例 5.14a　候选人 A 的哪些特点让你投他/她的票而不是投候选人 B 的票？

这样，问题本身就向回应者阐明了研究者希望他们谈一下候选人 A 的情况。如果所有回应者的回答都是以同样的参照标准进行的，研究者就可以对来自不同回应者的答案直接进行比较。

例 5.15　在你的居住地附近，你最喜欢的是什么？

例 5.15 中存在的不足：回答这样的问题，有些人可能只列举一点或者两点，而其他一些人则可能会列举很多。这种不同有可能反映出回应者观念或情感的重大差异。但是，研究清晰地表明，人们回答此类问题时所给出答案的数量与教育程度密切相关。另外，访谈员对答案的数量也有影响。

解决的方法是指明要列举的数目：

例 5.15a　请你指出在居住地周围你最喜欢的一个事物的特征。
例 5.15b　请你告诉我在你所居住地的周围你最喜欢的三样东西。

虽然这并不是解决所有类似问题的满意的方式，但这却是减少不同回应者的答案之间的不必要的差异的一个有效途径。

基本的一点是，由于回应者对正确回应的类型理解不同，导致

了答案的五花八门。因此，更好地说明所需答案的特征，会从根源上改变调查过程无信度的状况。

测量的类型/问题的类型

简 介

以上措施的设计，是为了最大限度地提高研究的信度，即保障人们在类似的环境中以最大程度的相似方式回答问题。不过，即使有极好的信度进行测量，可能仍然得不到所期望的测量结果。所给出的回答的真实性和表达了研究者需要或期望的表达的程度，叫作效度。这一节关于问题设计的一些方面的讨论，除了进一步提高问题的信度外，还能够提高调查测量的效度。

在这个问题的讨论上，有必要区分两类不同的问题：一类是用来测量事实或客观事件的问题；另一类是用来测量主观状态诸如态度、意见或情感的问题。尽管有些问题介于这两类问题之间，但是效度概念在客观性测量和主观性测量上还是有所不同的。

如果有可能通过一些独立的观察来检测答案的准确性，那么效度的测量与关于"真实"测量的意义相似。从理论上来说，人们可以得到独立的、准确的关于一年内某人使用 ATM 柜员机的次数。尽管实际上要得到这样的测量十分困难(例如：不可能获得相关记录)，但是，可以把效度理解为与客观情况一致。

相反，当人们被问到一些主观状态、情感、态度和意见时，除非回应者拥有接近其感受或意见的办法，否则，没有任何证实其答案的客观性方法。这种情况下，只能通过与其给出的另一些答案的比较，或者通过与其关乎测量内容的生活事实的比较，来确定人们主观状态报告的效度。对于这样的测量，没有真正独立直接测量的可能性，答案的意义必须从联系的方式中推断出来。

测量的不同水平

在社会科学中有 4 种不同的测量方法，对应就有 4 种不同的回

应者的任务，也就有 4 种不同的资料需要分析：

- 定类(nominal)：对人或事物进行无序的分类(如：你是男性还是女性)。
- 定序(ordinal)：按照同一标准对人或事进行有序分类(如：你怎样评价你的健康状况——非常好、好、一般还是不好)。
- 定距资料(interval data)：附加数字以便提供关于有序刺激之间或等级之间距离的有意义信息(事实上定距资料十分稀少，华氏温度是其中常见的例子之一)。
- 定比资料(ratio data)：数字除了被用来做值间间距之外，还用于表明各种数值间的比例。常见的例子是通过客观的、物理的量度表示的计数或者测量，例如：距离、重量或者压力(在你过上个生日时你多少岁)。

收集事实调查资料时，经常要求回应者把他们自己或自己的经历归入某一类，编制定类资料，或问他们一个数字(大多是定比资料)。"你是受雇佣的吗?"，"你结婚了吗?"，"你有关节炎吗?"，这些都是提供定类资料的例子；"你看过多少次病?"，"你的体重是多少?"，"你每小时的收入是多少?"，这些是要求回应者为定比资料提供真实数据的例子。

当收集事实资料时，也许会要求回应者做定序性的回答。例如，回应者被要求以相对大的类别来报告他们的收入，或者要求他们用非数值的言辞描述他们的行为(例如：经常、偶尔、极少或从不)。因为回应者在用定序术语报告事实时，研究者并不要求回答具有很高的精确性，或者认为确切地报告数据十分困难。但是，通常有一个真实的数字基础支撑着对某个事实问题的定序回答。

至于主观性资料的报告方面，情况则有所不同。尽管经过了多年的努力，尤其是心理物理学家(psycho-physical psychologist)的工作(例如：Thurstone & Chave，1929)，以求回应者用数字形式标明主观状态来实现定距和定比资料的设想，但是，大多数情况下，还是要求回应者提供关于主观状态的定类和定序资料。定类问题是"你的情感、意见或者感觉是属于哪一类的?"；定序问题是"你的情感、意见或者感觉处于这个连续体的哪个位置?"。

在设计调查工具时，研究者的一项基本工作是决定所期望测

量的种类。当做出这样的决定时，也就有了问题提问方式的一些清晰含义。

问题的类型

调查问题可以大致分为两类：那些为回应者提供好选项的问题（封闭式问题）和那些没有为回应者提供选项的问题（开放式问题）。

当目的是要把人们归于无序类别时（定类资料），研究者可以选择是提出封闭式问题还是开放式问题。实质上相同的问题可以用二者中任何一种形式设计。

开放式和封闭式问题的例子

例 5.16　你的健康状况如何？（开放式）

例 5.16a　你平常的健康状况是以下哪种情况？（读选项，封闭式）

例 5.17　你认为我们国家目前遇到的最重要的问题是什么？（开放式）

例 5.17a　这里有一个很多公民都关心的问题单。你认为目前国家遇到的最大问题是其中的哪项？（封闭式）

开放式问题有一些优点：首先，开放式问题使研究者可以得到意想不到的资料；其次，开放式问题更接近回应者的真实想法；第三，这点很重要，回应者喜欢有机会用自己的语言回答问题，如果只是从所提供的选项中选择答案，没有用自己的语言回答问题的机会，他们会很失望；最后，当可能的答案选项单长得难以向回应者提供时，用开放式问题更为适合。

除了不具备上述优点，封闭式问题通常是一种更令人满意的获得资料的方式。原因有以下三点：

- 当给出选项时，回应者能更可靠地完成问题的回答。
- 当给出选项时，研究者能更好地解释答案的意思（Schuman & Presser，1981）。
- 当提出一个完全开放式的问题时，很多人给出的是一些没有分析价值的相对稀少的答案。给回应者提供特定数量的

答案选项，足够多的人给出特定的值得分析的答案的可能性会提高。

最后，如果研究者希望得到定序资料，就要给回应者提供类别。如果在问题中没有可允许的答案排序方法的说明，回应者无法可靠地按照一个连续体将答案排序。当要求回应者完成一个定序任务时，对他们要完成的任务做进一步讨论是合理的，因为它可能是调查研究中最常见的一种测量类型。

表 5.1 给出了一个连续体（这个例子是让回应者给一些类分等级，但其一般方法适用于所有定序问题），其中研究者假定的范围是从最消极的感受到最积极的感受。研究者将回应者归入有序类别的方法，是在这个连续体上标上符号或标签。然后，要求回应者考虑这些标签，考虑他们自己的感受或意见，并将自己归入适当的类别。

关于从这些问题中获得的各种资料，有两点值得指出：第一，回应者可能对标签或者类型的意义有自己不同的理解。为了进行有意义的分析，最好的假定是：通常，那些报告自己感受"好"的人比那些报告自己感受"一般"的人感觉更积极。由于回应者对"好"和"一般"的理解程度和判别标准不同，导致测量缺少信度，但是测量仍然有意义（例如，与研究者希望测量的内在感觉状态相关）。第二，像这样的一个定序量表测量是相对的，人们选择特定标签或类型的分布状况，依赖于研究者所提供的特定量表。

再来考虑表 5.1 中的分等级量表和两种定序量表的编制方法。一种情况是，研究者使用的是三种描述选项量表：好、一般和不好。第二种情况是，研究者使用了五种描述选项：极其好、非常好、好、一般和不好。比较两种量表我们可以发现，后者增加的"极其好"、"非常好"决不会是简单地把"好"分成三部分，相反，它改变了量表的整体概念。人们除了对描述词作出回应外，还要为类别定序级别。当"一般"在量表中是第 4 点而不是第 2 点时，它更可能会被看作是连续体中的消极方面。因此，研究者希望更多回应者在 5 点量表中给出一个"好"的等级而不是在 3 点量表中。

表 5.1 对某事的感受

非常积极的				非常消极的
两类量表(two-category scale)				
好				不好
三类的量表(three-category scale)				
好		一般		不好
四类的量表(four-category scale)				
非常好	好	一般		不好
五类的量表(five-category scale)				
极其好(excellent)	非常好(very good)	好	一般	不好

如果能将这些量表用在恰当的地方(对人们排序)是十分有意义的。然而,就其自身而言,某个比例的总体觉得某事"好或更好"的一个陈述是不合理的,因为它意味着总体是在某种绝对意义上被描述的。事实上,如果问题不同,比例就会改变。当使用定序测量时,只有比较性陈述(或对关系的陈述)才是合理的:

- 在组群之间对同一问题的答案的比较(例如:A 组里认为候选人"好或者更好"的人数比 B 组多 20%)。
- 对可比样本的答案的跨时间比较(例如:在 1 月份认为候选人"好"的比 11 月份的人数多了 10%)。

相同的一般的评论适用于让回应者给选项排序而获得资料(例如:在学校、警察局和垃圾处理站中,你认为哪一个是最重要的城市服务设施?)。将一个选项排在首位或中间的百分数完全取决于所提供的特定项目单,当那些选项从根本上改变时,分布状况之间的比较就毫无意义了。

同意—不同意选项:一个特例

在调查研究中同意—不同意选项的使用十分频繁,因此要特别注意。在这些选项里,回应者的任务不同于他们将自己归属于某个类别顺序。最常用的方法就是将一个观点读给回应者听,然

后问他们是否同意那个观点。这个观点也是像表 5.1 所描述的那样被放置在连续体中的。通过指出他们的感受与观点的相似度（通过同意表现出来）或者同观点的背离度（通过不同意表现出来）来推测出回应者的反应在连续体中的位置。

当有人对同意—不同意形式的问题和直接分类形式的问题进行比较时，认为前者有很多缺点。比较下面两种问题形式的例子：

例 5.18　我的健康状况不好。你是极其同意、同意、不同意还是极其不同意呢？

例 5.18a　你是怎样评价你的健康状况的——极其好、非常好、好、一般或是不好？

第一种陈述形式有以下几个缺点：

- 等级量表把回应者分为五类，而同意—不同意问题的分析几乎总是把回应者分为两个组（同意的人或不同意的人）。因此，在分等级中可以得到更多的信息。
- 为了方便解释，同意—不同意问题只能对一个连续体的极端情况提问。如果陈述是“我的健康状况一般”，一个回应者会无法回答，因为健康状况只能或者是“好”或者是“不好”。这就限制了把人们排在连续体中间的能力。
- 回应者常常发现这是一种混乱：说他们的身体健康好本身就是不同意他们的身体不健康。
- 研究表明，当问题以这样的方式提出时，有些回应者特别喜欢说“同意”（或勉强同意）。就是说，如果问题 5.18 以否定的形式提出，那么有些人就会既同意他们的身体是“不健康”，也同意他们的身体“不是不健康”（Dillman & Tarnai，1991；Schuman & Presser，1981）。

对于一维测量任务来说，难以证明使用例5.18比使用例 5.18a 更合理。但是，它常用于获得对复杂陈述的回应。例如：

例 5.19　考虑到他们所处时代的经济条件的限制，有两个以上的孩子是不合理的。

这个问题同时问了三种事情：对经济状况的认识、对合理的最多孩子数的认识和对经济与家庭规模的关系的了解。

例子中存在的问题：如果一个人没有想到经济条件糟糕（这一点在问题中只是一个假设），也没有认识到经济条件与家庭规模的任何关系，他碰巧却认为一个家庭有两个孩子最合适，那么，他回答这个问题就很难。此外，不管一个人是说同意或者说不同意，都很难知道他同意或者不同意的是什么。

同意—不同意形式似乎是设计问题的一种相当简单的方法，而事实上，要用这种方式进行可靠、有用的测量是困难的，并且需要非常小心谨慎。通常，如果研究者避免使用同意—不同意的问题方式，就能得到更可靠、更有效和更容易阐明的资料。

事实报告效度的提高

当研究者提问关于回应者的一个事实问题时，目标是要回应者精确地予以报告。也就是说，研究者希望回应者给出的答案应该与研究者个人（假如研究者具有回答这个问题所需资料时）来回答这个问题的答案相同。目前有大量的关于事实材料报告的方法论文献资料。在各个领域，尤其在经济和健康领域，将口述事实报告与各种事实记录进行了比较研究（参看：Cannell, Marquis, & Laurent, 1977; Edwards et al., 1994; Edwards, Winn, & Collins, 1996）。

回应者很准确地回答了很多问题。例如，一个访谈报告了医院在 6 个月内 90％以上的夜间住院病例（Cannell et al., 1977）。但是，人们报告的质量是依赖于对他们提问的内容和方式。回应者报告的事件不太准确的基本原因有以下四个：

- 他们没有明白问题的含义。
- 他们不知道问题的答案。
- 他们理解问题，但是想不起来了。
- 他们不想在访谈中回答。

研究者可以通过一些措施克服这些潜在的问题。下面对这些

措施进行讨论。

理解问题

如果回应者对提出的问题有不同的理解，就肯定会产生误差。如前所述，当研究者试图量化具有复杂定义的事件时，诸如入室行窃次数或者看病次数，他们有两种选择：①给所有回应者提供定义；②让回应者提供对他们的经历进行详细的、复杂的分类所需要的信息，再用编码器对答案进行编码分类。

福勒(Fowler,1992)的研究表明，人们在回答包括含混不清术语的问题时，会产生一些失真的资料。研究者不能假设回应者不明白问题的意思时就会自动去询问研究者。为了尽可能地提高事实调查资料的效度，重要的、也是首要的一步，就是编写能被所有回应者理解的问题。

知识的缺乏

作为误差来源之一的知识缺乏有两种主要类型：①被抽样的回应者不知道问题的答案，但是在抽样的家庭里可能有其他家庭成员知道；②在抽样的家庭里没有人知道答案。解决第一种情况的方法就是正确选择回应者，而不是问题设计。通常一个家庭里的回应者被要求报告关于其他家庭成员的资料或者整个家庭的信息。解决办法包括以下几点：

- 确定并对最了解情况的家庭成员进行访谈。
- 使用的资料收集方法要给回应者向其他家庭成员咨询的机会。
- 取消回应者的代理权，回应者仅仅提供与他们自身相关的资料。

有时需要一种复杂的资料收集方法。例如，人口普查局进行的国家犯罪调查是由单身家庭成员提供信息以获得家庭犯罪的报告。但是，另外又直接向每个家庭里的成年人询问有关个人犯罪——诸如抢劫——的情况。如果访谈是以面对面的方式为基础进行的，给那些不在家的成员留下自填的问卷，或通过电话再进行第二次访谈，那么，与家庭其他成员的访谈成本就会减少。有一种

变通方式是让主要的回应者尽可能地报告所需的有关所有家庭成员的资料，然后给回应者寄出一个让其核实的概要，同时允许他与其他家庭成员商量。

当回应者答不出与自身有关的问题时，那就是问题设计上存在缺陷。从理论上来说，研究者能够区分开是回应者回忆不起来的信息还是他们根本就没有的信息。在任何情况下，研究者要做的是设计出几乎每个人都能回答的问题。可以进行以下选择：

- 修改问题，使询问的信息包含较少细节或容易回忆。
- 帮助回应者对答案进行评估。
- 改变或者放弃问题目标。

经常有一些问题要求答案的详细程度超过了研究目标的需要：

> 当研究目标是为了找出哪些人因为高血压而吃药时（一个比较简单的问题），提出的问题要求回应者说出他们使用的所有药物名称（一个很难回答的问题）。
>
> 当只需估计出收入的大概类别就能满足研究目的时，提出的问题却是开放式的（并且暗示要很详细）。

回忆会遵循几个很明显的原则：不很重要的小事比那些具有重要意义的事情更容易被遗忘；报告最近发生的事情比报告很久以前所发生的事情会好一些(Cannell，Marquis，& Laurent，1977)。有时可以通过提问更容易回忆的问题来提高报告的质量，这种改变问题目标的做法是很有意义的。例如，尽管期望回应者能报告过去一年发生的所有犯罪事件，但是如果研究者只要求报告最近6个月的，报告中的误差就会更小。

随着认知心理学家介入调查方法，导致了一系列新的问题设计策略的产生(Jabine，Straf，Tanur，& Tourangeau，1984；Sirken et al.，1999；Schwartz & Sudman，1996)。各种策略都致力于帮助回应者回忆所发生的事件（例如：通过提出可能的联想）或者把事情放到一定的时间背景中（例如：让回应者回忆发生在一年前的某些事）。研究表明，对于很多调查任务来说，回应者并不真正通过

回忆来回答问题，他们是在判断答案（例如：Burton & Blair，1991）。例如，如果要求回应者回答在某段时间内他们去食品店购买食品的次数，他们通常是根据一般情况去估计，而不是把每个事情都回忆起来。这种观察结论提醒研究者要为帮助回应者更好地做出判断设计问题方案。

最后，应该认识到，有些是研究者希望回应者回答但回应者回答不出的事情。例如，回应者并不了解由保险公司支付的医疗费用，如果真的想得到医疗支出的数目，就必须用直接从保险公司获得的资料来补充回应者的报告的数目（现金支付的）。

社会可接受性

在访谈中，有些事实或事件是回应者不愿意准确报告的。那些不合社会期望的健康情况诸如神经病和性病等明显地比其他健康情况更多地被隐瞒。对那些具有生命危险的住院经历，或者可能会给他们带来耻辱，或者使他们的生命受到威胁的经历，报告率较低（Cannell，Marquis，& Laurent，1977）。有关酒精消耗量的评估存在隐瞒不报的情况，这可能与回忆难度有关，但也与回应者对酗酒的社会道德顾忌有关。研究者发现，被逮捕和破产是另一种经常被少报而不可能忘记的事实（Locander，Sudman，& Bradburn，1976）。

在标准的访谈框架中可能会限制人们报告的内容。如果研究者希望有人承认一些很尴尬或非法的事情，那么就需要特别努力去劝说回应者，告诉他们所冒的风险极小和这个材料的重要性。当问到特别敏感的问题时，以下几点值得研究者考虑（也可参见：Catania，Gibson，Chitwood，& Coates，1990；Sudman & Bradburn，1982）：

1）使主观判断最小化，使精确性最大化。特别注意引导语和使用的词汇，因为在一些问题中，使用某些词汇会暗示出研究者对某些答案持否定态度。

在研究者与回应者进行交谈的过程中，研究者应该始终保持一种客观意识。如果有访谈员，那么他的提问

和举止会使回应者判断出研究者希望给出的答案。因此,研究者要特别注意回应者受到的语境提示对他们答案的影响。

2)使用自填问卷的资料收集措施。虽然资料不是结论性的,但是有一些证据表明电话访谈比个人访谈更易受社会可接受性倾向的影响(Aquilino,1994;de Leeuw & van de Zouwen,1988;Fowler,Roman,& Di,1988;Henson,Roth,& Cannell,1977;Mangione,Hingson,& Barret,1982)。在某些选项上,让回应者自填问卷形式比由访谈员提问可能受社会倾向性影响更少(例如:Aquilino,1994,1998;Aquilino & Losciuto,1990;Dillman & Tarnai,1991;Fowler,Roman,Di,1998;Hochstim,1967)。对于那些进行敏感话题的调查来说,应该考虑用邮件调查或者集体填答。个人访谈调查中也可以包括一些自填问卷形式的问题:只要让回应者在一个小册子里回答一系列的问题就可以了。如果调查使用了电脑辅助,回应者则直接将他们的答案输入电脑,也可以取得同样的效果。例如,在最近日益增加的关于非法使用毒品的报告中,这种方法显示了重大作用(Penne,Lessler,Beiler,& Caspar,1998;Tourangeau & Smith,1998)。最后,特纳、福塞斯和奥雷利(Turner,Forsyth,and O'Reilly,1998)指出,如果采用电话按键将答案直接输入电脑的办法取代访谈员访谈的方式,那么电话调查得到的与性和毒品有关的社会敏感活动的评估值的精确度更高。

3)机密性和匿名性。几乎所有的调查都向回应者保证处理他们的答案时坚持保密原则,非研究人员永远不会知道答案是由哪位回应者做出的。这些通常是在调查介绍中或者有预先的信件时,在信件中向回应者担保的。这些担保可以通过研究者签署的承诺书予以强化。没有标识码的自填问卷提供了一种不只是依靠信任保证答案匿名的方法。最后,如果调查是关于特别敏感的或涉及个人隐私的问题时,有一些详尽的调查策略,如随机回应

技术(random response techniques),这些策略保证了不会将回应者与他们提供的答案联系在一起(参见 Fox & Tracy,1986;Fowler,1995)。

另外,调查研究的一个局限值得重视,那就是回应者愿意告诉研究者的内容乃是在研究者设计的资料收集条件下的。如果没有特别的努力,有些问题可能不会在概率样本中问到。但是,本节讨论过一些措施,诸如努力为应答创设中立语境以及强调资料收集过程的精确性和客观性等,可能是最有益于问题的措施。任何问题,不管它看上去是多么好,都有可能使样本中的某些人感到尴尬。最好是在设计调查的各个阶段,使用各种方法以尽量减少对人们所作出的回答的社会压力。

描述主观状态答案效度的提高

如前所述,主观性问题的效度与客观性问题的效度有不同的意思。并不存在客观的评价标准,评估一个主观性测量的效度,只需要将答案在期望的方面与其他有关问题答案或者其他有关人格特点结合,并看其一致程度(参见 Turner & Martin,1984,其中有关于影响主观性测量效度问题的深入探讨)。

基本上只有三种措施可以提高主观性测量的效度:

1)尽可能地使问题具有可靠性。回顾前面讲到的问题的信度那些内容,处理表述的模糊性、标准化介绍、回应形式的模糊性,主观测量的效度随着主观性测量的不可靠程度提高而降低。尽量使问题对于所有的回应者都意味着同样的意义。由于主观性测量不是很可靠,它的效度就会降低。特别是定序测量表的信度问题,在主观状态的测量中,定序测量居于支配地位。提供的答案选项必须是一维的(例如:只涉及一个问题)和不变的(按顺序叙述,没有倒置)。

有疑问的量表

例 5.20　你是怎样评价你的工作的——非常有意义、有意义但有压力、不是很有意义但是没压力或一点意义都没有?

例 5.21　你是怎样评价你的工作的——非常有意义、有些意义、有意义或一点都没有意义?

问题 5.20 有两个测量特征即“有意义”和“压力”,它们不应该被放在一起,不是所有的选项都要列出。问题 5.20 如果既要问工作是否有意义又要问是否有压力,它就应该分成两个问题。在5.21中,可能有人会认为“有意义”(Rewarding)比“有些意义”(Somewhat rewarding)的意思更积极,并且产生疑惑不知道分类是怎样排序的。这些都是要尽力避免的常见问题。

2)如果在连续体中把人们按顺序分成几等,可能分类较多比分类较少要好。但是,回应者能够给出有序等级是有个区分精确度限制的,当类别的数量超出了回应者的辨别其感受的能力,大量的类别就容易产生不可靠性。并且,类别数量多也使得问题难以操作,尤其在电话访谈中更是如此。但是,从能测量出回应者之间的差异的角度来说,更多的类别意味着效度的增加。

3)提问复合型问题,用能测量出同一种主观状态的各种问题形式提问,然后把答案组合进一个量表中。所有问题的答案,既受到被测量的主观状态的影响,又受到回应者或问题特点的影响。有些回应者拒绝极端的分类,有的人倾向于说同意而较少说不同意。复合问题甚至还能帮助回应者作出独具风格的应答,并能提高测量过程的效度(Cronbach,1951;DeVellis,1991)。

有关主观测量意义的最重要一点即是他们的相对性。只有在刺激情况相同的时候才可进行分布状况的比较。在问题表述上的细微变化,答案选项数目的改变,甚至问卷中提问位置的改变,都会导致回答方式上的重大差异(参见:Schuman & Presser,1981;Sudman & Bradburn,1982;Turner & Martin,1984,其中有很多影响答案分布因素的例子)。不能对主观性问题的答案分布状况直接进行解释,只有将不同样本回答同一问题的答案进行比较,或对答案联系的方式进行研究后,对回应分布的解释才有意义。

问题设计和误差

社会调查的一个特点是问题的答案要充当测量的尺度。很明显,答案是否是好的测量尺度是衡量调查测量值质量的重要方面。问题可能会由于没有信度(产生不稳定的结果)或有偏误而成为不好的测量尺度,导致估计值会在某一方面背离客观事实(例如,少报酒后驾车被拘留的情况)。我们在怎样增加问题的可信度上所知甚少。本章所述的增加信度的方法可能是合理的,虽然可能会有其他要点被增加进来,但是编制清楚的问题以便给不同回应者提供一致的测量,这对一个好的测量来说总是具有建设性的一个步骤。

效度的问题更复杂。在某种意义上来说,每个被测量的变量都需要专门的研究来确定最合适的一套问题对其进行测量并形成最终测量有效性的估计值。本章所提到的很多有关改进报告的建议都来自一个长达20年的项目,这个项目是为了评估和提高与健康有关的变量的测量方法的(Cannell, Marquis, & Laurent, 1977; Cannell, Oksenberg, & Converse, 1977)。有很多领域需要更多有关效度的研究工作。

通过更好的设计问题来减少测量产生的误差是成本最低的提高调查评估的方法。对任何调查而言,关注问题设计和先行测试(在第6章将谈到),并充分利用已有的关于如何测量要测量的内容的研究文献,这都是十分重要的。另外,进一步形成一部评估测量效度并予以报告的文献显得十分必要。除了大量的工作尚待完成之外,鲁宾逊、谢弗、赖茨曼(Robinson, Shaver, and Wrightsman, 1997)、麦克道尔与纽厄尔(McDowell, and Newell, 1996)编辑了关于许多常用的多项选择测量效度的资料。其中,在论证尚须做的工作的同时,也证明了一些已有测量的有效性。

练　习

使用本章提到的标准来评估以下的问题是否是可靠的、可阐明的和有助于分析的测量，并试着编出更好的问题。

a. 测量收入：你赚多少钱？

b. 测量健康状况：你的健康状况如何？

c. 测量对生活的满意度：你是怎样评价你的生活的——非常好、比一般水平好、一般、勉强可以或者非常糟？

d. 测量对堕胎法的看法：请告诉我你是否同意以下的陈述：堕胎确实是一个严重的问题；除了在紧急情况下，堕胎应该是非法的？

编写一个关于两个变量的可能关系的假设（例如，良好的健康状况与高质量的保健有关，或者高质量的住房条件与高收入有关）。描述你所需的资料，以便分配给每人两个变量。然后起草一个问题（或一系列问题），用其描述在你的假设中被具体指定的两个变量的回应者，问题的答案可能会提供你所需的信息。指明你的问题是要求得到事实的还是主观的信息，并且你得到的资料是否有定类、定序、定距还是定比的属性。



6

评估调查问题和调查工具

Evaluating Survey Questions and Instruments

设计一个令人满意的调查工具包括几个方面工作:选择能够满足研究目的的问题;对问题进行测试以确保提问者和回答者都能按计划提问和回答;将问题制作成调查表以最大程度地方便答问和访谈。本章将描述设计令人满意的调查工具的几个步骤。

每一个调查都要有访谈计划或者调查问卷，前者是为访谈员编写的脚本，后者由被调查者自己阅读并填写。这些问卷形式或者计算机程序形式的文献，就是通常所说的调查工具。

通过第 5 章的讨论，弄清了什么样的问题才是一个好问题以及怎样把问题作为测量工具。这是设计令人满意的调查工具的基础。但是，还需要采取一系列切实可行的步骤来制一份令人满意的数据收集工具。萨德曼和布拉德伯恩(Sudman & Bradburn, 1982)、康弗斯和普雷瑟(Converse & Presser, 1986)、布拉德伯恩和萨德曼(Bradburn & Sudman, 1992)以及福勒(Fowler, 1995)对此做了非常详细具体的探讨。

设计调查工具包括两个方面：确定要测量什么、设计并测试问题。第一步通常是界定调查目标，当然这些目标可以根据随后的问题测试而修订。然后就是选择和测试问题。调查工具开发过程包括以下几个步骤：

- 重点小组讨论。
- 草拟一系列尝试性的问题。
- 批判性地复查问题以找出不足。
- 独特的实验室访谈(不要套用设想的资料收集程序)。
- 把问题组合成调查工具。
- 运用近似于设想的数据收集程序进行前测。

界定调查目标

要设计一份令人满意的调查工具，其先决条件是确定要测量的内容。这看上去很简单且不证自明，但这一步骤经常被忽视，从而影响调查结果。首先，写出调查应该达成的目标，这一步非常有价值。在设计调查工具时，研究者常抵挡不住诱惑，要添加一些相关但于研究目的无益的问题。抵制这种诱惑的方法，是对调查目的作出明确表述，并以此保证某一特定调查领域的结论能够测量出来。其次，研究者应当列出为达成研究目的所要测量的内容清单。内容清单所列出的不是问题，而是要测量的变量，并且要按照

有利于形成观点的类别或方面列出。

除了要列出测量的变量之外，还应制定一个分析计划。可以假设，样本设计是调查的良好开端。研究者必须全面考虑，要特别评估人群中的那些子群体。因此，研究者要提炼这些观点以明确列出：①哪些变量设计为因变量，评估这些因变量的集中趋势（如均值或值的分布）；②哪些变量设计为自变量，以理解分布状况和联系模式；③哪些变量是控制变量或插入变量，以解释观察所得的模式并检验不同的假设。

总之，目的的陈述、测量变量种类的清单和分析计划草案，这三份文件是开发调查工具过程中必不可缺的组成部分。

初步的问题设计步骤

焦点小组(focus group)

拟定结构化的系列问题之前，有必要就所要研究的问题与研究总体中的人进行焦点式讨论。讨论的根本目的是对回应者的实际与研究目标所内含的抽象概念进行比较。

例如，研究目的是测量看医生的次数，群组讨论就集中于讨论怎样才算是看医生。这里两个关键概念是“看”(visit)和“医生”(doctor)。可以询问参与者与医生的各种交往（如电话咨询、照X光或实验室测试、接种疫苗），并问他们是否认为这些就是“看医生”。也可以询问他们所接触的关涉他们健康的各种人员（如心理学者、精神病医师、医生助理、眼科医师、验光员、临床医学家）的情况，以及是否认为这些人也是医生。

仅仅这个讨论就可提供给我们至少三种信息：

- 可以算作是“看医生”的交往类型。这一信息可以帮助研究者改进研究目的和提问的措辞，以明确应当或不应当包括某些内容。例如，要不要包括电话咨询呢？如果医生办公室有见习护士，算不算医生呢？
- 人们知道些什么。例如：是不是每个人都清楚精神病医师

有医学博士学位而心理学家却没有呢？关于医疗保健工作者的背景、培训和资格，人们是怎样认识和理解的呢？

- 对一些关键词语或术语的理解。“医生”是指获得医学博士的人呢？还是指更普通身着白大褂、提供健康服务的专业人员一类的人呢？

焦点小组讨论最好6～8人参加，围绕关系调查要测量内容的人们的理解、经验和感受等进行讨论。小组讨论人数到底多少为宜，这可能不一样。但是，在调查工具开发的早期阶段进行几次焦点小组讨论，将最终有益于调查工具的制定。

拟定问题

拥有了所要测量内容的清单，研究者就尝试去寻找一个或一系列问题来构建测量清单上那些变量的工具。诸如有关背景或人口统计之类的很多问题，对很多调查来说都是常规性的问题。审视一下芝加哥大学民意研究中心做的“一般社会调查”问卷是有裨益的。另外，还可以通过密歇根大学“高校间政治和社会研究协会”(ICPSR)在线获得很多调查。各主要调查机构的调查工具原件复印件也是有用的参考资料。研究者可以从这些资料中获悉如何表述详细而精确的问题，如何生成标准化问题，以及如何格式化调查工具。

借鉴利用他人已做工作是非常明智的。当然，最好是复查那些以前做过这个研究课题的研究者用过的问题。另外，如果一些问题已经用于提问了其他样本，收集可比资料能增加研究的普适性。但是，有人曾经用过某个问题这一事实并不能保证它就是一个好问题，或者就是一个适合于某个给定调查的问题。由于研究者不加批判地使用，导致了很多有缺陷的问题被反复提问。对所有问题都应当加以测试，以确保它们适用于特定研究的总体、背景和目标。

前调查评估

批判性的系统复查

一旦拟定好一系列问题，紧接着就是对它们进行批判性、有系统的复查。莱斯勒和福塞斯（Lessler & Forsyth，1996）对此以问题形式列出了要注意的事项。福勒（Fowler，1995）也提出了一系列在测试前能运用的问题的标准。尽管这些清单都不是很全面的，但都识别出了一些问题特征，能标示不良问题。运用这些序列有助于识别需要修订的问题，还能标记出在以后测试阶段需要引起注意的问题。

认知实验室访谈

一旦问题拟定好、经过批判性地复查及修改之后，下一步就是确认它们是不是人们都能理解和回答的问题。焦点小组讨论有助于研究者洞察理解问题，但未能提供一个评估问题表述精确性或回应难易程度的平台。在设计问题的早期阶段，研究者也可以通过朋友、亲戚和同事来试验设计的问题，从中也可以有所收获。大多数调查工具的早期版本都包含一些含糊不清的问题，这些问题读起来难懂，实际上任何人都无法回答。

问题拟定好后，在进行正式现场前测之前，还要进行一次正式的测试，这就是人们通常所说的认知测试（DeMaio & Rothgeb，1996；Forsyth & Lessler，1992；Fowler，1995；Lessler & Tourangeau，1989；Willis，DeMaio，& Harris-Kojetin，1999）。虽然认知访谈形式各异，但它们确实有一些共同特点。首先，接受访谈的是一些志愿者，为了帮助研究者理解问题是如何起作用的，他们花费的时间比资料收集本身实际所需时间多得多。通常情况下，受访者都有报酬，他们置身于实验室环境，访谈可以录下来。

这些访谈通常不是一般访谈员能做的。有时是由认知心理学家来做访谈员，有时是由调查者本人或高级访谈指导员来做的。

在大多数情况下，访谈员需要对每个问题的目标都十分清楚，以便能觉察出受访者理解问题或形成答案的方式中所出现的问题。

一份典型性的草案要求向受访者提供一系列准备好的问题，然后以某种方式收集他们理解问题和回答问题方式的信息。有时，在受访者准备答案时，要求他们"大声地想"(think aloud)；有时会问受访者一些有关他们怎样理解及回答问题方式的问题。最常见的两种任务是：

- 要求受访者用自己的话说出他们是如何看待所提问题的。
- 要求受访者解释他们是如何从其他答案中选出某个答案的。

问题的核心是要得到有关受访者理解和准备答案的过程的足够信息，这样才能评估他们是否以研究者所希望的方式来做的。大多数认知测试要回答以下四种问题：

- 对问题的理解是否一致？
- 受访者是否拥有回答问题所需的信息？
- 答案是否准确地表达了受访者的真正意思？
- 答案是否能有效测量设计问题时所要测量的内容？

从实验室访谈所得的资料是有局限的。首先，这种方式的访谈通常很少(常少于 10 次)，因为这种工作是劳动密集型的，并且在大多数机构中，只能由很少一部分人主持。其次，访谈是在人为的环境中进行的，是由样本中的志愿者能够并乐意承担的任务，样本的一个不规则的部分可能难以完成这个任务。尽管如此，人们越来越认为这种访谈是设计和评价调查工具的关键一步。在实验室环境中无法得到一致理解或回答的问题，当然在现实调查中也难以很好地运行(Royston，1989)。现场前测阶段识别出的理解和答问上的困难并不如在实验室访谈阶段所获得的可靠，因为后者可以检测答问过程。实验室认知访谈法大多用来测试访谈草案，但理解和答问方面的问题与困难同样存在于自填问卷的调查中。正如后面将要阐述的，尽管对自填问卷形式进行标准测试通常包括与认知访谈所用相似的询问性问题，但是受访者的理解方式在以口头问答方式进行访谈的过程中更加明显。因此，要测试为自填问卷式访谈设计的问题，使用一个口头认知访谈可以有效识别标准前测不能识别的问题。

调查工具的设计、格式和编排

一旦问题准备就绪付诸最后前测时，就有必要把问题放进一个表格以方便访谈员或自我填答使用。第一步就是给问题排序。很多研究者都喜欢将相对比较简单、直截了当的问题放在最前面，以吸引受访者。通常，需要认真思考的或人们认为很敏感的话题放在调查工具的中间或后面。一个实用的好办法就是按节给问题编号：A1、A2、B1、B2 等，这样，增加或删除问题后就不必再为每个问题编号了。

不管调查是采用访谈员访谈还是自我填答的方式，问卷的编排和格式化都是为了尽可能地便利访谈员或回应者。对于访谈员访谈的调查工具，以下规则有助于达成这一目的：

- 采用区分对待访谈员读给受访者的话语与指导语的约定。常见的约定是指导语用大写字母，要读出的话语用小写字母。
- 如果访谈采用纸笔方式而没有电脑辅助，忽略那些不适合特定访谈对象的问题，在指导语中对此做出明示。当然，电脑辅助的访谈可以自动跳题。
- 用括号表示可供选择的词语，如“(他/她)”或“(丈夫/妻子)”等习惯用法。如果访谈员看到括号内的提示，就容易处理了。另外一个习惯做法就是，将词语全部大写(如，SPOUSE)，这时访谈员必须提供问题中没有的词语。电脑辅助的访谈能提供供选择的词语，而不是由访谈员自己根据情况调整措辞。
- 检查以确保访谈员要说的话都写下来了，不仅包括叙述问题的词，还包括过渡词、对问题的介绍、必要的概念和解释。

这些基本原则也同样适用于自我填答问卷式调查，即主要目的是使问卷用起来更便利。如果有什么区别的话，那就是自我填答问卷调查的格式化更重要。与访谈员相反，受访者没经过训练，他们常常没有动机来做好问卷，而且也不是按处理问卷的能力来

选取的。自我填答问卷调查有如下五个指导原则：

- 自我填答问卷总体上是无需另外说明的。不需要阅读指导语，因为它们并不是一直读出来的。
- 自我填答问卷应主要限定为封闭式答案。受访者只需勾画复选框、选择答案或者圈数字就行了。如果要求受访者用自己的话回答问题，答案常常是不完整、模糊而难以编码的，因而作为测量的价值是极其有限的。
- 自我填答问卷的问题类型数量应少。在一份调查问卷中，如果受访者完成同种类型的任务或回答同种类型的问题越多，他们就越不太可能感到迷惑；同时，他们做起来也就越容易。
- 问卷的排版应看上去清晰、整洁。缩小照片(photo reduction，或其他将很多问题排在同一页上的策略)，与把同样多的问题分布在更多页面上相比，会降低受访者做出反应的速率。
- 通过书面和可视线索为受访者提供更多信息，反复提示如何继续。如果人们不清楚自己被要求做什么，他们肯定就会出现混淆，因此要尽力使每一点都简单清晰。

电脑辅助测量工具的设计还在继续发展中。毫无疑问，上述很多原则都同样适用于电脑辅助的测量工具。但是，库珀、汉森和萨达斯基(Couper，Hansen，& Sadowsky，1997)发现访谈员在使用一些复杂的 CAPI 程序时难度很大，这暗示还需要进一步发展(同样可参见 Couper，1999)。迪尔曼(Dillman，2000)描述了他设计的最佳互联网调查工具的成果，这项工作同样也在进展中。

现场前测

一旦调查工具设计完成，研究者认为基本上可以使用了，这时就应该对调查工具和调查程序进行现场前测。进行这种前测的目的是找出资料收集草案和调查工具如何在真实环境下运行。

对访谈计划前测

由权威调查机构做过的传统的前测，其通常做法是，从与要调查的总体相同或者相似的总体中抽取受访者，由经验丰富的访谈员对这些受访者做 20～50 个访谈。访谈员在这种前测中扮演两种角色：一方面，他们是访谈员，执行访谈任务；另一方面，他们又是资料收集过程的观察者，要向研究者报告调查程序和调查工具中任何有待改进之处。尽管有时访谈员是个别地进行汇报，但这种反馈可能最为典型的是在集体汇报会上进行。

上述前测是调查设计过程的重要组成部分，它特别重要的一个功能就是从访谈员的角度来测验测试工具的可用性，包括问题和排版。但是，这种测试也有一些局限性。访谈员用以判定存在问题的标准通常不很明确，并且几乎可以肯定地说，访谈员们在认为什么是问题上也观点不一。另外，群组讨论在收集有关前测经验的系统性信息方面也不是一个完美的办法。

为使前测经验更加系统和更有价值，研究者补充了一些步骤。创举之一就是要求访谈员除了群组汇报之外，还要给每个问题填写一个简明的分级评分表格，在表格中要求访谈员评价每个问题是不是：①读与写同样方便；②受访者理解问题的一致性；③受访者能准确回答问题(Fowler，1995)。很明显，访谈员不得不猜测受访者是否理解了问题并能准确回答，但是，无论如何他们都会这样做。表格的优点在于，它要求访谈员除了关注惯常注意的调查工具的实用性方面之外，还要全面系统关注问题设计的上述方面。另外，让访谈员对这些问题分级，可以让调查者能概括访谈员的报告，用更一致的方式识别所设计的问题中存在的不足。

就现场前测而言，一种更重要且更有用的创新，是使用磁带录音和行为编码来评价调查问题。如果受访者同意(这几乎不成问题)，就可轻而易举地录下前测访谈，不管是面对面还是通过电话进行的。然后，训练有素的编码者通过听录音并用一致的方式评价问答过程中出现的问题。

业已证明，有三种行为在识别调查问题中存在疑点时是最为重要的(Fowler & Cannell，1996；Okesenberg，Cannell，& Kalton，

1991)：①访谈员是否按照所写读出了问题；②受访者是否要求解释说明；③受访者最初给出的答复是不是不充分的而需要访谈员再去探查。研究发现，在访谈中，问题总是导致或不导致这些行为的产生，也就是说，有些问题访谈员一直误读，从而导致受访者要求解释或者一致地回答不充分。这个编码过程不能识别受访者不能一致理解的所有问题，但是，任何一种行为在前测访谈中出现率达15%及以上时，就表明某一问题极有可能会引发不正常的资料，或者明显受访谈员的影响(Fowler，1991；Fowler & Mangione，1990)。

前测访谈中行为编码的另外一个好处是其结果具有系统性，而且可以复制。因此问题评估过程是不受研究者和访谈员的主观意见影响的，并且可以形成对不充分问题进行判断的具体的、可重复使用的标准。

从计算机辅助访谈的前测中可以得到痕迹文件(Trace files)，这是第三种信息来源。当访谈有计算机辅助时，有可以重新得到访谈员敲击的关键键。这些文件可以识别访谈员在何处返回到前面的屏幕或前面的问题。返回到前面的问题会减缓访谈进程，也通常表明了那个问题的流畅性没有设计好。查看"帮助"功能的使用情况，就可以确定什么地方需要帮助以及各种帮助功能到底有什么作用。同时，检视痕迹文件还可以查看结果是否有系统性，是否可以计量(Couper，Hansen，& Sadowsky，1997)。

自我填答式问卷的前测

与访谈调查工具相比，自我填答问卷调查工具应更多做些前测，原因是访谈员可以解决一些研究者在设计调查工具时没有解决的问题。不幸的是，对自我填答测量工具进行前测也有一定难度，因为理解上的问题和答问时的困难不很明显。尽管有人运用观察人们如何填表或如何与计算机交互作用作为识别模糊的或令人迷惑的问题和指导语的方法，但这不如通过录音和访谈的行为编码那样令人满意。

或许对自我填答问卷进行前测的最好方法就是亲自与一组可能的受访者交流。如果是基于电脑的调查，受访者可以使用便携

式电脑。首先，受访者应就像自己是调查的一部分一样完成调查问卷，然后研究者可以组织关于调查工具的讨论。第一个话题当然是指导语是否明确，第二个话题是问题是否明晰，第三个话题是在理解期望的答案或按预期回答问题方面是否有困难。

除了团体测试之外，基于计算机的调查工具的可用性往往还得益于一对一的测试。在这种测试中，观察受访者与计算机及问题间的交互作用。

调试计算机辅助调查工具

让访谈员或受访者测试调查工具，这可以提供关于使用方便的信息，但不能提供资料收集草案是否正确的信息。需要关注的重点是“跳过”指令。

计算机辅助的一个巨大优势是帮助受访者和访谈员正确测定偶发事件。什么时候该问哪些问题或者如何表达，与对前面问题的回答紧密相关。当然，要保证“跳过”指令的准确性就要对问卷调查工具的最后版本进行仔细校对。但是，检查计算机辅助调查工具的准确性所面临的挑战远远大于检查问卷调查工具。其难点在于测试者看不出跳过了哪些问题，因此他们可能忽略某一问题本应问及却被略过这一事实。校对打印出的程序并做大范围的测试是很重要的步骤。但是，如果调查工具很长而且还包含复杂的可能事件，仅做这些就不够了。因此，调查一旦开始，标准做法就是把前期所得答案的分布状态列成表格。只有通过检查这些结果，研究者才能确信关于可能事件的指令在按计划运行。

调查工具的长度

一个令人满意的前测可以发现完成一个调查工具要花多长时间。判定访谈长度的标准应该包括成本、回应速率的影响以及受访者的能力限制和答题意愿。自我填答问卷的长度影响成本和回应速率，其程度随调查人群和调查主题不同而有所差异，对此很难做出概括。另外，能对受访者访谈多长时间也很难做出概括。

如果研究者发现要问的问题多于能问的问题，那么有两种选择：首先，研究者完全可以删除一些问题；其次，研究者也可以针对不同的受访者提不同层次的问题。后一种方法增加了调查的复杂性，降低了对变量估计的精确度，但比将问题删除要好一些。计算机辅助资料收集的一个明显优势就是，借助它可以方便地把这种设计完成。

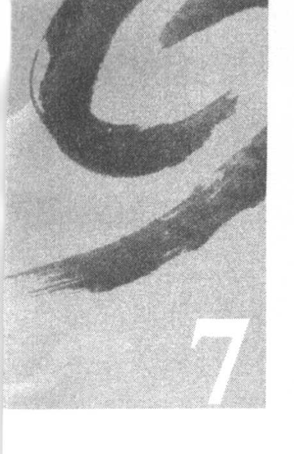
总　结

人们曾经认为，对问题的评价很大程度上是个主观过程，随访谈员和研究者的偏好而具有不确定性。现在我们知道，我们可以避免这种主观性。对所有接受访谈的人来说，调查问题的意思都应该一样；而回答调查问题也应当是绝大多数或几乎全部接受访谈的人都能做的；访谈计划用语的描述应当全面，访谈员可以据此开展访谈。

显然，无论问题表述的多么清楚，总有些接受访谈的人会遇到困难，而且总有些访谈员会误读。如果问题必须修改，在修改前对问题的不足要做出一些判断。设计和评价调查工具的关键是收集有关理解、答问以及访谈员和受访者如何使用草案的信息。这样就能判断是否需要修改调查问题和调查工具。在真正调查前，对调查问题和调查工具做令人满意的评价，是使调查获得预期效果的关键。这是减少调查评估误差的最省事的方法之一。尽管还需要确定怎样评价调查问题才算最经济有效，但上述程序构成了一组实用技术，如果采用，将会对提高调查资料质量产生重大而积极的影响。

练　习

运用第5章练习中设计的问题，把它们转换成访谈员能以标准化方式控制的问题系列。根据需要进行前测并修正。再把同样的问题转换成自填问卷调查形式，并进行前测。



访谈调查

Survey Interviewing

访谈员从三个方面影响调查估计值:对访谈得到的回应率起主要作用;负责培训和激励受访者;以标准化、不带偏见的方式主持访谈和问答过程。本章讨论选择、训练和监督访谈员以及给予访谈员的一系列措施对于最大限度地减少调查中人为误差的意义。



访谈员工作概述

尽管有些调查采用自我填答方式，但让访谈员提问并记录答案却仍是调查测量过程的常见组成部分，其中既包括面对面访谈也包括电话访谈。由于访谈员在资料收集过程中的核心作用，他们就极有可能影响所收集资料的质量。管理访谈员是一项困难的工作，对于个人访谈员工作的研究尤其如此。本章的目标是让读者明白访谈员应该做什么、管理访谈员的恰当程序以及访谈员的操作行为对调查估计值的影响。

访谈员在调查资料收集过程中扮演三个主要角色：

- 选定受访者并谋取合作。
- 培训并激励受访者做好访谈工作。
- 提问、记录答案，探查不完整的答案以确保答案符合提问的目的。

获得合作

访谈员要与受访者取得联系以获得合作。这一工作的难度因样本差异而不同。要让受访者愿意回答问题时随时能够找到访谈员，访谈员要能够与难以接触的受访者取得并保持联系，并且在面对面访谈中，访谈员要能够并愿意去受访者所在地。

尽管很多样本个体乐意接受访谈，但要得到事先未通知到或最初不太愿参与访谈的人的合作，这是访谈员必须做的最困难却是最重要的工作之一。访谈员在这个问题上比在任何其他问题上更容易失败。

无疑有些访谈员比其他访谈员在争取受访者的合作方面更加出色。同时可清楚地看到，不同的个人风格也会产生不同的影响。有些高效率的访谈员办事很有条理，而有的则风度翩翩。经验告诉我们，那些在获取受访者合作方面很出色的访谈员有两种共同特点：第一，他们都非常自信。他们在陈述研究项目时似乎完全相信，受访者肯定会合作。他们说话的语调和内容都表明他们相信

访谈肯定会成功。第二,他们有迅速调动人兴趣的诀窍,因此交往能关注和适应个体特点。也许这太过于以任务为取向,但却适应了个体的需要、顾虑和环境。用预先设计好的草稿照本宣科不是一种获得合作的有效方法。

虽然这些访谈员技巧对所有调查都十分重要,但电话调查却对这些技巧提出了挑战。因为在电话调查中受访者事先没有得到通知(如随机电话拨号访谈),或者对调查主题不感兴趣。

培训和激励受访者

受访者的表现,诸如报告的精确性等,与他们对访谈的定向有关。业已证明,访谈员对受访者的目标设定起了重要作用(Cannell & Fowler, 1964; Cannell, Oksenberg, & Converse, 1977; Fowler & Mangione, 1990)。例如,访谈工作草率的访谈员会鼓励受访者快点回答问题。而慢慢读出问题的访谈员则以非言语方式暗示受访者可以多花点时间以作出经过审慎思考的、准确的答案,结果他们的确能得到更准确的回答。研究也表明,访谈员鼓励受访者的方式会影响受访者对自己应当做什么及做得如何的感知(Cannell, Groves, Magilvey et al. , 1987; Cannell, Okesenberg, & Converse, 1977; Fowler & Mangione, 1990; Marquis, Cannell, & Laurent, 1972)。

毫无疑问,大多数受访者对自己应当做什么和如何做都知之甚少。访谈员明确或含蓄地教会了受访者该怎样做,这一点常不受重视,但却是访谈员工作中十分重要的部分。

成为标准化访谈员

调查研究者倾向于认为,答案中的差异是由于受访者所说内容(即他们的观点和经历)不同,而不是由于他们所受刺激(即提问的措辞、提问的背景及提问的方式)不同。绝大多数访谈员的培训目的,都是使受训者成为一个不影响所获答案的标准化了的访谈员。研究者试图使之标准化的访谈员行为有五种:呈示研究和任务的方式;提问的方式;探讨不完整答案(即没有达到问题目的的答案)的方式;记录答案的方式;访谈处理人际关系的方式。下面

对这五个方面略做进一步陈述。

1)呈现研究。受访者应对研究目的有一共同理解,因为目的性与他们答题的方式有关。对诸如保密性、自愿性以及谁运用结果等问题的假设都会影响到答案。优秀的访谈员会让所有受访者对访谈项目做相似的取向,这样访谈背景就能保持一致。

2)提问。调查问题应当按所写的那样准确提出,不做任何变异或措辞改变。有时,即使问题用词的细小改变都会对问题的回答产生重要的影响。

3)探查。如果受访者没有完整回答问题,访谈员必须提出一些补充问题以引出更好的答案,这就是探查。访谈员应当间接地探查不完整的问题,即尽量不让答案出现重合的情况。如果调查工具设计得好,包括重复性提问诸如"还有什么吗?"、"能多告诉我一些吗?"、"你的意思是什么?"的标准化探查可以应付绝大多数情况。

4)记录答案。记录答案应标准化,以便在这个阶段没有访谈员导致的变异。提出开放式问题时,访谈员要逐字记答案,也就是说要准确记录下受访者所用的词语,而不阐释、概括或者遗漏。在封闭式问题中,给了受访者选项,访谈员应该在受访者真正选定后才记下答案。如果访谈员将受访者的话编码成他们并未选择的类型,就可能导致前后矛盾。

5)人际关系。访谈中的人际关系也要以标准化方式处理。访谈员不可避免地会把某些人口统计学特征带入访谈,如性别、年龄及教育情况等。但是,强调交互关系的专业方面并集中注意于研究任务,就会最大程度地减小关系中的个人方面。一般地,应教导访谈员不要讲述他们自己的事情或者表达与访谈主题内容有关的个人意见和看法。访谈员不能就受访者做出的答案有任何评价。总之,要避免访谈员个人的、特殊的行为特征的影响,因为这些特征在访谈员之间是不同的。表现得像专业人士而不像朋友,这有助于实现访谈员与受访者间关系的标准化。没有证据表明友善的人际关系风格会从本质上提高报告的准确度,而且它很可能对报告的精确度有负面的影响(Fowler & Mangione,1990)。

如果访谈员和受访者的社会背景不同,情况就复杂得多了。

在这种情况下，交流就不如背景相似时那样随意和轻松。有证据表明，那些采取措施来缓冲这种复杂性（如交流时加入一些幽默）的访谈员可能得到更有效的访谈（Fowler & Mangione，1990）。但是，努力使受访者放松不应该偏离基本的专业化交往，目标应放在取得令人满意的结果上。

访谈员工作的意义

现在我们可以清楚地看到，访谈是一项困难的工作。此外，这一工作做不好就会导致三种不同的调查资料错误：

- 如果访谈员不努力取得受访者的合作，样本就会失真且会导致偏见。
- 如果访谈员影响资料的方式不一致，调查估计值的准确性会降低，围绕估计值的误差会增多。
- 如果访谈员不能正确训练和激励受访者，或者没能建立报告所需的恰当的人际环境，答案整体上会不准确或有偏见。

由于这些情况都可能导致错误，所以，应该提倡研究人员使用优秀的访谈员。很多方面都会影响访谈员的工作质量：征募和选拔、培训、监督、设计令人满意的问题以及运用有效的程序。下面五部分就将讨论这些方面对访谈员工作的影响。

访谈员的征募和选拔

在访谈员工作规范中就访谈员的一些无关调查资料质量的特征做出了规定：

1）访谈员要具备良好的读写技能。不说是绝大多数也是很多访谈员都借助电脑工作，因此访谈员还要会打字并基本熟悉电脑。绝大多数调查研究机构要求访谈员有中学毕业文凭，而且很多都要求或偏向有大学学历的访谈员。

2）访谈大多是兼职工作，很难做到每周工作 40 小时来做基本人口调查，调查机构对访谈员的需求总是时多时少。因此，可能做访谈员的人必须是能容忍收入的不连续性的人，或者是作为找到

更稳定工作前的人。访谈员的报酬对于一个大学毕业生来说通常是不高的,而且经常没有福利,如健康保险等。访谈调查人员很难仅靠这份工作的收入来维持长久的生活。

3)个人家庭访谈员在时间上必须有弹性,调查要求访谈员在受访者有空时就到场。电话访谈的优点之一就是独立的访谈员能够按照可预测的轮班时间工作,尽管晚上和周末是几乎所有一般总体调查的黄金时间。

4)个人家庭访谈员必须具有机动性,这就排除了那些身体不好或者没有车的人。但对电话访谈员来说,所有这些限制都不重要。

除了这些实际性工作要求以外,还没有足够的研究基础能证明对访谈员候选人有选择偏好。例如,经验丰富的访谈员更容易得到受访者的合作,原因仅仅是因为那些难以获得受访者合作的人已经被淘汰了,但是,没有记录表明经验会对资料质量有积极影响。有迹象证明,时间长了,访谈员会粗心大意,收集到的资料更难令人满意(Bradburn, Sudman, & Associates, 1979; Cannell, Marquis, & Laurent, 1977; Fowler & Mangione, 1990)。

同样地,也没有必要要求访谈员拥有关于研究主题的专业知识。事实上,当受访者没有表达清楚的时候,知识渊博的访谈员会自认为知道受访者在说什么,他们会比没接受过这个领域训练的访谈员更多地把自己的意见加到受访者的回答中。因此,除非访谈观察或报告要求访谈员必须拥有渊博的知识背景,通常没有专业知识背景但受过训练的访谈员才是最佳人选。

尽管有一些从不规范的样本中获得的证据表明女性访谈员更好一些,但访谈员的年龄、教育和性别很少与资料质量有关(Fowler & Mangione, 1990; Groves, 1989)。总地来说,研究者总会选派最好的访谈员采访受访者,而不要考虑其人口统计学的特征。例外的是调查主题直接针对民族或宗教(或任何其他人口统计学特征),或者针对受访者对同一或不同群体的人的感情。例如,如果要采访人们的反闪族感受,访谈员的犹太身份就会导致回答的差异性(Robinson & Rhode, 1946)。同样,黑人和白人也会因访谈员的肤色而对种族问题作出不同的回答(Schuman & Converse,

1971)。

但是,要特别注意的是,种族匹配并不一定能提高报告的质量。有两项研究发现,黑人受访者接受白人访谈员关于福利收入(Weiss,1968)和投票(Anderson,Silver,& Abramson,1998)问题的访谈时,比接受黑人访谈员访谈时表达更准确。

毫无疑问,研究者应该考虑调查主题与访谈员和受访者的人口统计学特征之间的互动关系。如果种族划分(或其他特点)对将给出的答案极为重要,那么就需要考虑控制访谈员与受访者特征之间的关系,以便测量访谈员对资料的影响(Groves,1989)。但对于大多数调查来说,控制访谈任务所存在的实际困难和成本,以及缺乏对影响的可预测性,都不支持控制访谈员和受访谈员的人口统计学特征。

最后,志愿访谈人员几乎总是不能很好地完成概率抽样调查,失败原因有如下几个方面:第一,很难要求志愿者参加冗长的训练,因而他们没有经过系统的训练;第二,很难不收自愿但素质不高的访谈员,因而调查回应率通常不高;另外,志愿访谈员耗损率通常很高。

上述讨论为研究者在选择访谈员时提供了指导性原则。在某些专业化的情况下,访谈员的民族背景、年龄或者性别会影响答案。例如,十几岁的青少年在接受年纪大的女性访谈员的访谈时表现会有很大差异(Erilich & Riesman,1961)。但是,在绝大多数调查中,是由特定的研究工作需求支配着访谈员的组合。由于背景或个性特征而排除某些人做访谈员,这是没有根据的。相反,要组建优秀的访谈员群体,关键是要给予他们良好培训和精心监督。另外,由于无法事先识别优秀的访谈员,事后裁减掉能力较差的访谈员,对于建立素质优良的访谈员队伍来说是关键而必要的。

访谈员的培训

访谈员可以接受培训的经验种类很多。很明显,投入培训的时间长短、培训课程的类型及课程内容都取决于特定组织环境以

及访谈员将要做的工作。另外,对于在初始培训尤其最初的现场经验培训、继续学习及开始工作后的再培训等方面投入多大的精力,人们并未达成共识。尽管如此,所有关注资料质量的专业调查机构对新招募的访谈员都会进行至少某些类型(通常是面对面)的培训。下面概括了合理的访谈员培训应当包括的项目。

培训内容

培训内容既包括适用于所有调查工作的一般知识,也包括专门针对特定访谈工作的知识。要讨论的共同话题包括如下几个方面:

- 联系受访者并介绍研究的那些程序。
- 在设计调查工具时所用的有关问题表述和跳答指令的那些协约,这些协约能使访谈员一贯地按标准提问。
- 间接地探查不完整答案的那些程序。
- 记录开放式问题和封闭式问题答案的那些程序。
- 不带偏见地处理人际交往问题的那些原则和指南。
- 如何使用计算机辅助访谈程序。

另外,很多研究机构认为,让访谈员感受如何使访谈适合于整个研究过程是个不错的主意。为此,他们常试着让访谈员熟悉抽样过程、编码、分析的类型以及调查结果的报告。这类信息有助于访谈员回答受访者提出的问题,并且还可激励访谈员并帮助他们理解访谈工作。

具体到某一特定的课题,访谈员还应知道以下几点:

- 课题的具体目的,包括发起、研究的总体目标以及研究的预期运用。这些信息是为受访者提供正确的问题答案并取得他们合作的基础。
- 抽样的具体方法,也是为回答受访者提出的问题提供基础。此外,还需要一些关于如何实施基本样本设计的培训。
- 关于特定问题目的的细节——不一定是在分析中的作用,但至少是要获得的信息种类。
- 关于保密以及给予受访者何种恰当承诺的具体措施。

培训过程

培训访谈员有六种基本方法：书面资料、讲座和展示、基于电脑的培训指南、计划练习、角色扮演及观察早期访谈。书面资料通常有两类，首先，最好有一本全面的访谈员手册，全面描述访谈过程；另外，对于每一特定的研究，都应该有一些针对特定项目的说明。最好是当面培训访谈员，并在本地开展课题以节省书面材料的准备工作。但新培训的访谈员却认为，培训期间他们有很多很多的资料和信息要理解。把程序写下来可以让访谈员更悠闲地复习，也能进一步清晰、准确地陈述学习内容。

显然，不管是只培训一个还是培训一群访谈员，演讲和展示在任何访谈员培训中都起作用。除了总体展示必需程序和技能外，大多数培训者都发现，展示标准化访谈是让访谈员明白怎样开展访谈的快捷而有效的方式。讲座大多使用录像带作补充，见习访谈或其他访谈活动的录像带是培训访谈员的有效工具。还可以购买普通培训磁带用来培训访谈员。

计算机辅助访谈的广泛使用意味着在访谈员培训过程中必须向访谈员教授如何使用以电脑为基础的测量工具。使用最为广泛的调查系统都有可以整合进一般访谈培训中去的基于电脑的指导。

由于这些是新技术，结构化的、有督导的练习是访谈员训练中最重要的组成部分之一。让访谈员轮流扮演受访者和访谈员的角色，这是很常用的方法。练习应包括争取合作及处理问答的过程。对那些由访谈员不认识的真实的受访者参与的访谈实践进行监控，这也是很有价值的。对于个人访谈，监督者可以陪同或观察新访谈员练习访谈，或者检查访谈的磁带录音。在电话访谈中，可以直接监控访谈，或者先录音，供以后检查。

有两个研究（Billiet & Loosveldt，1988；Fowler & Mangione，1990）表明，如果培训时间不足一天，则不能培训出令人满意的访谈员，访谈员不能按指示完成任务，并对调查资料会有不利影响。专业调查机构的培训课程一般持续 2～5 天。培训时间长短取决于多个因素，其中包括要培训的访谈员总人数和培训内容的复杂

程度。但是，培训质量的关键很可能是监督下的访谈练习的数量。

监 督

有效监督的关键是获得评价访谈员表现的信息，并投入评价信息和提供及时反馈所需的时间和资源。监督访谈员的表现主要有四个方面：成本、回应率、已完成问卷的质量以及访谈质量。运用中央控制设备来监控电话访谈比其他访谈要容易得多。

成 本

监控访谈员的成本需要及时了解时间花费、工作效率（通常是完成的访谈）以及访谈用车的开支。高成本的电话访谈员可能工作效率低下、被拒绝率高（一次拒绝所花时间与一次访谈所花时间几乎差不多），或者找理由（如编辑访谈、削铅笔）每小时只打几个电话。高成本的个人家庭访谈员可能住得距离样本所在地较远，为调查跑路太少，或访谈时间不合适（周末和晚上显然最有效率），或者回应率太低。

回应率

访谈员适时监控回应率（特别是拒绝率）是十分重要的，但这也不容易做到。主要有以下三个问题：

- 对于个人访谈而不是计算机中央控制的电话访谈来说，要及时掌握访谈员的结果信息是较难的。
- 访谈员把失败的访谈归入其他类别，这不能准确显示拒绝率。
- 面对面的访谈中访谈员之间获得的回应率没有可比性，这样每个访谈员的被拒绝率不总能预示访谈成绩。这一问题在中央控制的电话访谈中就不存在。

调查结束前是无法准确计算出回应率的。但是，访谈员在资料收集过程中努力识别拒绝者可以提醒监督员注意这些问题，也是访谈监督的一个重要部分。很难帮助那些有回应率困难的访谈

员。在电话调查中，监督员可以听介绍，并在访谈（或非访谈）结束后立即就访谈员如何提高效率提供反馈意见。对于面对面的访谈来说，监督工作更困难，因为除非监督员与访谈员一起去，否则他无法观察到访谈员所运用的访谈方法。因此，监督员通常只能满足于听访谈员所做的样本简介。

监督员可以给访谈员提供有益的暗示。保证访谈员对调查有全面了解十分重要，让访谈员练习简明清晰地回答一般性问题是十分有益的。除了将介绍工作做细致外，监督员可能需要强调访谈者对与之打交道的人们和调查项目及其价值的总体感受。但是，再培训的作用也是很有限的，总有些人终究不能得到令人满意的回应率。尽管这是个压力，但是，有一个保持高回应率的最有效办法就是从研究中剔除低效的访谈员。

调查工具完成后的复查

如果访谈员使用纸笔调查工具，就应当对完成了的调查问卷样本进行复查，以便评估访谈员收集到的资料的质量。在复查一个已完成的调查时，显然能够查看记录是否清晰、跳答指令运用是否得当、所得答案是否完整以便编码等。另外，查看一个已完成的访谈，通过比较记录摘要或阐释，能很好地看到访谈员在多大程度上逐字记录了回应者的答案。对计算机辅助访谈而言，这些问题（除了与陈述答案相关联的记录和探查之外）是不相关的。

问答过程

通过复查已完成的调查工具无法监控访谈质量，复查已完成的调查工具并不能告诉监督员有关访谈员开展访谈工作的方式，也不能告诉监督员那些答案是如何得来的。监督员只有直接观察访谈过程，才可能了解到这些。

中央控制的电话调查允许直接监控访谈员收集数据的方式，监督员能够也应该随时监控访谈员。监督员应该全面系统地听取每个访谈员所做访谈样本的全部或一部分，从而评价（还包括其他内容）研究介绍语的恰当性、按照所写内容提问的严格性、探查的适当性和间接性、处理访谈中的人际关系的恰当性等。如果监督

员对访谈员的诸如此类的各方面工作能够全面地、制度化地进行评价，这一过程会非常有效(Cannell & Okesenberg，1988)。

如果访谈员是在受访者家里或其他较远的地方进行调查，就更难监督问答过程了。只有两种解决办法：监督员以观察员身份陪同访谈员，或者对访谈进行录音。如果不录音或观察，研究者就无法评估访谈质量。所有测量过程中最重要的方面都没有监控，不能辨别出不合格的访谈员以便对其进行再培训，研究者除了说已经告诉访谈员该做什么之外无法报告访谈的质量。实际上，从访谈员的角度看，如果标准化访谈只是作为培训的重点却没有进一步深入实际，就很难让人相信它的重要性。福勒和曼吉温(Fowler and Mangione，1990)提供的证据表明，如果不用磁带录音直接监控个人访谈员，他们就不太可能以培训所学方式进行访谈。福勒和曼吉温(Fowler and Mangione，1990)以及比里特和洛夫特(Billet and Loosveldt，1998)都发现，如果以这种方式直接监控访谈员，资料质量就能提高。现在很清楚，对访谈程序进行直接监控应当是管理有素的调查的一部分。

调查问题

尽管培训和监督对令人满意的访谈来说很重要，但研究者能采取的最好措施是设计出良好的调查工具。研究表明，某些问题总是被误读，有些问题的回答总是不完整而需要访谈员去探查以得到完整的答案(Fowler，1991；Fowler & Cannell，1996；Fowler & Mangione，1990；Oksenberg et al.，1991)。运用第 6 章阐述过的前测方法可以鉴别这些问题。

访谈员要探查、解释或澄清的越多，他们对答案的影响也就越大。调查工具越完善，访谈员就越有可能开展标准化的、令人满意的访谈。福勒和曼吉温(Fowler and Mangione，1990)及福勒(Fowler，1991)详细讨论良好的问题设计在令人满意的访谈中的作用。

访谈过程

培训和激励回应者

研究证明，良好的问题设计在促进访谈标准化方面有重要作用(Cannell et al.，1987；Cannell，Oksenberg & Converse，1977；Miller & Cannell，1977)。例如，研究者可以帮助访谈员用一致的方式培训受访者，访谈开始前，访谈员可以读出下列内容：

> 由于大多数人还没有接受过这种调查，所以在开始调查前，我需要先给你说一些有关访谈过程的事情。调查中将问你两种问题，一种问题需要请你用自己的话来回答问题，这时我会记下你所说的每一个字，不做任何概括；另一种问题，我们会提供一些答案，请你从中选出与你的想法最接近的选项，尽管其中的任何答案都不可能与你的想法完全一样，但还是请你选出一个最接近你想法的选项，以便我们可以比较你与其他人的答案。

有趣的是，访谈员很喜欢用这样的介绍，因为它介绍了受访者要做的事，使问答过程更顺利。实际上，优秀的访谈员能够自己做出这样的介绍。提供清楚的介绍，让所有访谈员做同样的事，这样可以减少差异。另外，这种介绍对访谈员的成绩也有积极影响。一旦访谈员读出这些介绍，清楚解释了访谈要达到的目标，访谈工作就易于按照要求方式进行了，并且不易出错，因为此时受访者也知道访谈员被期望做什么了(Fowler & Mangione，1990)。

给受访者的标准化介绍也可用以确定操作的目标和标准：

> 你尽可能准确地回答问题是非常重要的，不必匆忙。如果必要时你可以查阅有关记录。如果有什么问题不清楚，可以要求我进一步解释。

这样的表述可以保证受访者对自己优先考虑的事情有共同理解。有的访谈员无意识地向受访者允诺，只要他们参加访谈，将会是很轻松容易的事，那些匆忙的访谈员向接受访谈的人传递了一种速度比准确度重要的信息。如果读出的是上面讨论的这种介绍，就会促使访谈员和受访者都以追求准确性和资料的质量为努力方向，访谈员之间的差异降低了，访谈员与受访者双方良好操作的可能性却提高了。

坎内尔、沃克森伯格和康弗斯(Cannell, Oksenberg, and Converse, 1977)报告了一种更好的方法。在允许受访者参与访谈前，要求他们签字承诺尽力给出准确而完整的回答。预期会有很多人拒绝，但并未发生。尽管回应率没有受这种形式影响，但是报告的质量却提高了。在电话访谈中用口头策略进行同样操作，也证明能有效提高报告质量(Cannell et al. ,1987)。

坎内尔还试图将访谈员给予受访者的强化进行标准化。访谈员经常会无意中强化不期望受访者出现的行为(如快而不假思索的答案。Marquis, Cannell, & Laurent, 1972)。坎内尔、沃克森伯格和康弗斯(Cannell, Oksenberg, & Convers ,1977)报告，如果设计的访谈计划迫使访谈员赞赏受访者的良好行为(如检查记录或慢慢答问)，受访者的报告质量就会提高。平常运用这些程序有些难度，但这项工作强调的是把访谈员做的不恰当的强化降到最低程度。

总之，除了直接的问答程序，访谈员的工作还有其他重要成分。特别是访谈员要负责告诉受访者访谈的进程：受访者应该做什么、访谈员要做什么、双方的共同目的是什么。这个方面的工作主要是留给访谈员自己做，不同的访谈员会以影响资料的不同方式来做这些事情，就不足为奇了。通过为受访者开发标准化的介绍程序，研究者可以使访谈员的工作更容易，减少访谈员之间的差异，改善访谈双方的行为，使访谈过程更为完善。

标准化表述

前面提到过，严格按照写出的问题提问是标准化测量的基础，但并非每个人都同意这个观点(Tanur, 1991)。标准化访谈的批评

者注意到,并非所有受访者对某些问题的理解都是一致的。他们认为,在这种情况下,让访谈员自由澄清或解释问题将提高资料质量(例如,Schober & Conrad,1997)。同时批评者指出,如果访谈员只用预先写好的表述时,有些资料收集任务(如,从几个不同的人或者事中收集相同的资料)会歪曲人际交互关系。批评者认为,在这一事例中,如果给予访谈员话语上更大的灵活性,就可以使访谈双方的交互关系更加愉快(Schaeffer,1992)。

对标准化访谈的批评中,有些主要是针对问题设计低劣所造成的结果(参见 Suchman & Jordan,1990)。如果问题表述不清,或者给访谈员的脚本不方便使用,常用的补救方法是写出更好的问题,而不是让访谈员重新设计问题(Beatty,1995)。有确凿证据证实,如果访谈员可以灵活自如地重述或解释问题,他们常会改变问题原意并进而使资料质量更糟而不是更好(Fowler & Mangione,1990)。但是,如果给予访谈员更大的灵活性,确实有些问题(例如,重复,或本来只需给少数受访者详细解释的却对所有的受访者都讲述,这会令人讨厌)能得到更好的解决。另外,如果访谈员改变问题表述,也并未一致证明会增加与访谈员有关的误差或回应误差(Dykema,Lepkowski,& Blixt,1997;Fowler & Mangione,1990)。

有些实验是关于给访谈员更多提问和探查答案的处置权的(Conrad & Schober,2000;Schober & Conrad,1997)。迄今为止,实验结果也不统一:有些报告的准确度可能提高了,却大大增加了对访谈员的培训量及访谈时间。何时、如何给予访谈员更多的灵活性还有待进一步实验。同时,对于大多数调查来说,设计出访谈员能够且愿意严格按所写形式提问的问题,仍然是开展令人满意的调查的主要方法。

访谈的真实性

访谈员可能杜撰访谈是个值得关注的问题。这种事情发生的可能性随样本、访谈人员和实际访谈过程不同而不同。在很大程

度上，人们关注的是访谈员在受访者家里或在自己家里做电话访谈时的真实性问题。在这种情况下，监督员不能观察到资料收集的实际过程。做访谈所花费的时间之多，足以刺激一个访谈员杜撰一个访谈而不花时间和精力去真正做访谈。

从长远来看，预防访谈做假的最好办法，也许是建立一支对研究质量和研究机构负责的访谈员队伍。新雇佣的访谈员最有可能做出这种虚假事情，但是，即使是拥有经验丰富的专业化工作人员的机构，也需经常对访谈样本做定期检查，以确保访谈得以真正实施。

解决真实性问题有两种办法：其一，给所有的受访者寄一封简短的事后问卷，询问他们对访谈的反应。其二，也是更常用的办法，要求访谈员从每个受访者那里抄回电话，然后监督员抽样打电话检查。事先让访谈员知道会通过邮件或电话检查访谈的真实性，这样对那些搞欺骗的访谈员有威慑作用。顺便指出，只有能够说访谈资料是经过核实了的，才可以保证资料使用者放心使用这些资料。

访谈过程对调查误差的影响

本章开始就提到，访谈员会影响回应率、报告的准确度以及测量的一致性或精确度。这些在评估调查质量中都起核心作用。

令人满意的调查运作一个最明显的作用就是对回应率的影响。尽管第3章已详细讨论过这个问题，但仍需重申的是，访谈人员的素质对任何调查所得回应率都是至关重要的。

要测量访谈员在问答过程中带来的误差是更为困难的，其调查误差通常难以发现。当对主观状态提问时，对偏见或不准确性进行客观检查通常没有意义，这在第5章已经谈过。但在有些研究中，研究者有对要求受访人报告的事实的客观测量，允许对报告的精确度进行评估。在一个研究中(Cannell, Marquis, & Laurent, 1977)，受访家庭样本是那些一年前家里有人住院的家庭。通过比较住院访谈报告和医院病历可以评价报告的准确度，即只要看报

告的住院比例就可测量报告的准确性。

在这个研究中，每个访谈员访谈的人数与访谈中没报告的住院百分比的相关性很高($r=0.72$)。访谈人数多的访谈员，不管他们承受什么样的压力，都比访谈人数少的访谈员所收集的准确数据少很多。

另一个研究运用同样的标准(报告的住院比例；Cannell & Fowler，1964)，也得出了相似的结论。在本研究中，让访谈员的受访者中一半一次性地报告住院情况，而让另一半在访谈员结束其他健康访谈离开后自我填答关于住院情况的表格。研究发现，受访者报告的住院准确率高的访谈员，其受访者在访谈员走后做自我填答式问卷时也做得令人满意($r=0.65$)。这一研究结果表明，不仅访谈员在影响受访者报告失误中起重要作用，而且访谈员影响受访者表现的另一个方面是访谈员在何种程度上激励受访者做出正确反应。在这两个研究中，访谈员对报告准确度的影响是明显的。

在没有有效资料的情况下，无法评估准确度，但是，也可以评价访谈员影响受访者回答问题的程度。如果所有访谈员都以非常标准化的方式做访谈，就无法通过知道访谈员是谁而解释答案中的任何变异。如果从知道谁做的访谈可以部分地预测答案的话，那么就可以说访谈员是不恰当地影响了答案。格罗夫斯(Groves，1989)全面讨论了计算访谈员影响问题答案程度的方法，并对计算访谈员影响的研究结果做了概括。结果表明，对于访谈员所问的很多问题，并不能看出访谈员对答案有任何影响。但在大多数调查中，访谈员严重的影响了1/3到1/2的问题答案。

这些访谈员影响的结果是增加了调查估计值周围的标准误差。增值的大小由类内相关系数的大小及访谈员任务数量的平均值决定(参看 Groves，1989；Kish，1962)。如果类内相关系数是0.01(Groves 发现这是平均值)，每个访谈员平均做31次访谈，均值的标准误差就会比仅从样本设计估算得到的数值高14%。当访谈员的平均访谈数接近50，组内相关数是0.02时，标准误差的估算值会增加41%。

从这些讨论中，可以得出几点关于访谈员在调查资料总误差

结构中的作用的结论如下：

- 除了作用于回应率之外，访谈员还关系到受访者在调查过程中给出不准确答案的程度以及测量的不一致性。现有证据明确显示，访谈员是造成多种误差的一个重要原因。
- 访谈员接受的培训和监督能极大提高访谈员的一致性，从而提高估计值的可靠性并降低偏见。特别是那些接受培训最少（例如，少于一天）或得到关于访谈质量反馈最少或没有反馈的访谈员，是较差的访谈员。
- 建构受访者培训和指导的程序，把访谈员不合理的反馈降到最低，并较多控制访谈员行为。这些措施能减少访谈员对调查资料的影响，并提高总体准确度。
- 良好的问题设计是获得令人满意的访谈的关键。
- 一个未被重视的设计选项是访谈员平均任务量。尽管访谈员用得少，其培训和管理成本也会更低，但如果一个研究者让个体访谈员承担大量的访谈任务，那么他就要为数据的可靠性付出代价。减少访谈员的平均工作量，从而提高调查估计值的精确度，这是最有成本效益的方法。
- 几乎所有关于调查估计值可靠性的报告都忽视了访谈员对资料的影响。这部分是由于以便利性和地理邻近性等非随机方式给访谈员布置任务时，研究人员无法从抽样影响中分离出访谈员的影响。但是，对于大多数调查项目来说，访谈员的影响是误差的重要来源，任何忽视访谈员影响的关于调查估计值精确性的报告，都会低估调查误差。

总之，访谈员在调查资料误差中的作用未被普遍重视。尽管绝大多数调查研究者都知道访谈员需要培训，但是培训程序和访谈员监督相差很大，且常不充分。除培训和监督之外，在尽可能地降低访谈员影响方面，研究者往往没有付出努力。但是，调查设计的这些方面是提高调查资料质量最有成本效益的方法，因此在调查设计和研究中应当充分重视访谈员对调查估计值的影响。

练　习

用录音带录下你或其他人使用标准访谈计划时所进行的角色模仿访谈(第6章中提出的问题,或其他渠道的计划)。然后听录音,通过记下每个问题所出现的下列错误来系统地评价访谈员的表现:提问没有完全按所写的进行;探查不完整答案时用带偏见(直接的)方式;没有探查不明确的答案;或其他任何可能带偏见或不标准的人际交往行为。如果是以集体方式进行评价,就特别有指导意义,这样可以讨论访谈员误差。



8 整理调查资料

Preparing Survey Data for Analysis

调查答案通常都会被转变成资料文档格式以便使用电脑进行分析。本章将讨论资料格式的选择和使用、代码开发、编码过程和管理、资料输入以及检查资料过程。



通常调查收集得到的资料总需要转变为适合电脑分析处理的格式。本章主要讨论将完成的问卷和调查访谈转变为电脑能够阅读和处理的格式的过程。编码或资料归并的过程牵涉到以下五个独立的阶段：

- 确定使用的格式（将资料组织成文档的方式）。
- 设计代码（将回应者的答案编辑为电脑能处理的规则）。
- 编码（将回答转换成标准类别的过程）。
- 资料输入（将资料转变为机读格式）。
- 资料清理（在开始资料分析前，对资料文档的准确性、完整性和一致性进行最后的检查）。

将答案输入资料文档中可能会出现两种错误：第一，在人们记录答案或数据时，随时可能出现抄写错误。第二，也可能是编码决策错误，误用了答案与代码值匹配的规则。控制质量的方法与特定的资料输入和所选的编码过程有关。接下来就将讨论这些问题。

格式化资料文档

每个分析程序包都有它自己的资料格式化规则。在设计资料输入过程前最重要的一步，是决定使用哪些程序进行资料分析以及用这些程序能处理哪些有关文档格式和漏缺资料的具体规则。“记录”这个术语在这里是指与个体、案例或访谈有关的所有资料。一个记录能包括一个或多个系列的资料。在历史上，曾用一张 80 栏的卡片作为标准单位来记录资料，它相当于一个系列的资料。现在则是把资料直接写入硬盘或软盘，通常仍然是将每个记录或访谈的内容以有序的形式储存。尽管协议和规则会因使用不同的设备和程序而有所变化，但是仍然有如下一些普遍使用的形式：

1）即使没有使用真正的资料卡，用一张目录卡来指定资料位置的形式仍然很常用。在这种格式中，每写满 80 栏，就开始新的一个系列。但是，很多分析程序可以很轻松地处理比用 80 栏的卡片记录的资料更长的资料，如果使用了一个这样的程序包，就没必

要受一个目录卡格式的限制。

2)在每一个记录或者调查中,为每个回应者设定的一系列标识码应出现在每个资料系列的同一位置。在每一个系列或者卡的同一位置有一个系列号或者卡号也有助于检查数据文档的完整性。如果要归类资料和检查文档的完整性,这些标识符就会保持资料的顺序。

3)如果按照资料在调查工具中出现的顺序对其进行编码的话,就能使编码、资料输入和编程工作更为简单。这能减少误差的出现,并且也是一种无须成本的控制资料质量的方法。

4)有些程序也可以接受在单个的空间或栏目中的多重编码,有些则不能。最好是为每个包含资料的空间只设定一个确定的入口。同样,有些计算机程序将空白翻译为0,而另外一些程序没有这样翻译。如果0是有所指,最好是编码成0而不是留下空白,如果旨在为无回应者编码,就应该使用一些特定的值(通常是数字形式而不是空白)。

创建代码

代码就是把答案转化为数字或者把数字转化为答案的一套规则(一些系统接受按字母形式的值,但是调查中的大多数代码只使用数字形式)。哪个数字对应哪个答案的情况与电脑无关,但是它对资料进行可靠编码和正确翻译十分重要,代码有时可能会模糊不清。因此,指定数字代表答案(或其他结果)应遵循一些规则。另外,代码设计可以用来最大程度地减小编码和分析时的误差。以下是一些常用的编制代码的原则:

1)要确定有缺失资料的代码,这些代码之间会有以下几方面的区别:

- 不确认的资料,即那些由于访谈员或者回应者的不良表现没能获得可编码的资料;一些研究者也倾向于用不同的代码来区分回应者拒绝回答的问题和没有回答的问题。
- 不适用的资料,即对某些特定人群不适用的资料(例如:向

没有住院的人询问住院时间)。

- “不知道”的答案会被归入不确定或作为缺失资料的一种独特类型。

2)指定标识符时要保持一致性,对“不确定”,“不知道”或“其他”的回答总是使用前后一致的代码。代码越具有一致性,译码员和程序设计员所产生的误差就越少。

3)尽可能让代码与现实中的数字一致,代码标号要准确(例如:把 45 周岁编码为 45)。同时,如果没有特别的原因,最好以答案在调查中出现的顺序进行数字排序。

如果提供给回应者的是选项答案或答案形式是高度结构化的,那么,编码员的工作就只是为给定的一系列答案指定数码并说明缺失资料。但是,如果回应者被要求用自己的语言回答问题,那么答案范围不可能完全提前预测。对这种开放式回答的问题,代码的开发就是研究者赖以鉴别答案类别并给答案排序的过程。

代码开发的理念是,为具有相似分析意义的分组答案建立类别,区别对待不同的答案。如果分类太细,其结果就是许多类别里只有很少的资料,并且难以分析和编码;另一方面,大的、过于笼统的分类则会掩盖答案间重要的差异。衡量好的代码的一个标准就是给每个答案明确惟一的标识符,另外一个标准就是把答案归入有分析意义的类别中。在有明确的分析计划的情况下,才能评估后一个标准的执行情况。

为了建立这样的代码,必须注意:

- 明确答案的哪些特点具有分析意义。重要的第一步是,草记每个问题的答案间的各种差异,这些差异在研究者看来应是重要的。
- 把早期访谈得到的答案制成表格。然后编制一个代码草稿,以便给这些答案分类。
- 把这个分类方案用于 10～20 个其他的访谈中,必要时则予以修改。
- 给选项为“其他”的答案进行单独的编码,并让编码员对这些答案做出注释。这些注释既可以用来扩充和阐明代码或增加所需的类型,也提供了包括在“其他”类型中的答案记录。

- 这样的注释也应该方便编码员与研究者进行交流以解决编码规则中的问题和模糊的地方，然后研究者针对有问题的地方进行修改。

这些步骤与有效的编码检查（下面将提到的）一起，会产生详尽的没有重复的类别系统。这一系统可以明确地将每个答案放进一个固定位置，是编码员、编码监督者以及研究者可以共同使用的。

编码和资料输入的方法

从产生自动制表到20世纪70年代末或80年代初，这段时期里，问题回答和资料输入方法包括三个步骤：一是把答案记录在纸上；二是训练有素的编码员将答案转化为代码数字并将数字写在专门的编码簿中；三是打孔操作员将编码的数字打到80栏的IBM卡上。

即使事实上从没依照这些步骤操作，但是了解用这些措施减小误差进行质量控制的过程是有指导意义的。

1)访谈员记录：没有切实有效的方法来检查访谈员是否准确地记录了答案(例如：正确勾选了复选框)。然而，将访谈员作出编码决定的可能性最小化却是一种很好的尝试。对于开放型的答案首先需要逐字记录，然后再由训练有素、接受监督的编码员进行编码。当遇到根本不清楚怎么确定分类的时候，应该要求访谈员随时记录下所有相关的资料，这样就可以在编码时经常检阅和处理这些资料。

2)编码：编码过程的质量控制包括下列步骤：

- 培训编码员，包括让所有的编码员对几个相同的调查资料进行编码，然后比较结果，以确定他们都是以同样的方式进行编码的。
- 对每个编码员的工作进行单独的抽样，以检查编码。这样做可以达到两个目的：①辨别出哪些编码员做出错误的编码决策；②找出模糊的、需要说明的编码规则。

- 采取措施让编码员对不知道怎么进行编码的答案写出注释。监督员应定时评估这些作为编码检查系统扩展部分的注释，这有助于监督者找出需要注意的编码员或代码。

3)资料输入：新的电脑技术给资料输入过程带来了巨大变化。前面提到的旧技术是使用打孔机器，它通过在 IBM 卡上打孔记录资料。这种机器功能单一、价格也比较昂贵，因此只有专门化的打孔设备才可能使用这样的技术。由于电脑相对廉价和功能多样，它无处不在，并且任何电脑都可以用来输入资料。因此，不用再将资料送到专门的打孔场所，就可以由任何人在任何地方输入资料。此外，电脑可以进行程序编辑通过以下方面改进资料输入过程：

- 在特定空间只允许输入合法代码。
- 能检查资料以确保它们与以前输入的资料一致。
- 能自动确保恰当处理意外出现的问题（即当一系列问题只向部分回应者提出时，能够对意外出现的指令编程，这样跳过的问题空间会被自动填上合适的代码）。

尽管这些检查没有识别那些符合编程规则的资料输入的误差，但是很多的资料输入误差可以在发现的时候轻松地进行改正。

伴随这种新技术的出现，目前产生了两种比旧的三步输入法更为常用的资料输入法。第一种是两步法，通常又称之为直接资料输入法：当访谈员或回应者填好了书面形式的调查工具，编码和资料输入工作同步进行。通过让第二个人独立进行编码和资料输入来检查这个编码和资料输入的过程。

值得指出的另一个两步过程法是光学扫描法。现今扫描技术的发展日新月异，扫描技术用于资料输入有两个方式：

一是对回答选项可以进行条形编码，通过扫描仪扫描所选答案旁的条形码以输入数值。这种方法的优点在于，它使那些没有资料输入技能的人也可以做资料输入工作。

二是扫描诸如标准化考试之类的特定答题纸或者表格。这种已经存在了好多年的做法可使资料输入的成本较低。其花费只在所需的扫描仪的费用，即使对于有特定目的的表格，也不过就是再增加设置表格和印刷表格的费用。

从历史发展来看，光学扫描仪在调查中的使用率呈下降趋势，

这是因为：

- 使用表格不方便，调查研究者希望问卷尽可能地便于使用。
- 为相对小型的调查建立特别的表格成本比较高。
- 可能导致严重的资料丢失，当要求未调动起积极性或未受过训练的回应者使用这种方法时尤其如此。

用手工检查缺失项目以识别机器读不出的符号的方法，可以解决最后一个问题。但是，要在调查中广泛使用光学扫描仪，可能要以提高技术为前提。与以前的扫描仪相比，现代化的扫描仪更能兼容不完善的符号。只要使它们适应于方便的调查工具，现代化扫描仪能够用于扫描各种版本的资料。

尽管技术进步到了编码员可以扫描手写答案供后续编码之用，但只有对固定选项和提前编码的资料扫描时才能顺利进行。鉴于更好的设备仍然相对昂贵，扫描仪在资料输入中的作用将会继续增大。迪尔曼和米勒（Dillman & Miller，1998）、迪尔曼（Dillman，2000）以及布洛姆和赖伯格（Blom & Lyberg，1998）很好地总结了现行扫描方法及其局限性。

比旧三步法更常用的第二种方法是一步法，它又被称为电脑辅助电话访谈法（CATI）、电脑辅助个人访谈法（CAPI）或者电脑辅助自填问卷法（CASI）。它要求访谈员或者回应者直接把答案输入电脑，然后进行所需的编码。访谈中不会使用纸和笔。从资料输入的角度看，一步法与在英特网上收集资料的方法本质上是相同的，在这两种方法中，回答问题和资料输入是同步进行的。

正如在第 4 章所述，电脑辅助资料收集方法越来越普及。对于电话访谈来说，每个访谈员需在电信局有个终端。终端屏幕上显示出问题，访谈员要读出问题，回应者回答，然后访谈员将与答案相对应的数值输入终端，而数值的输入会引发下一个问题出现在屏幕上。研究员可以给电脑编程，让它只接受合法的输入并检查输入的答案是否与以前输入资料的一致，这样访谈员能够清楚出现在回应者答案里明显不一致的地方。手提电脑或其他可以携带的个人计算机也为家庭访谈提供了同样的资料收集方式，这与用网站进行资料收集的方式是一样的。

所有使用电脑辅助资料收集系统，都有以下几个优点：

- 电脑可以处理复杂的问题类型，而在纸笔调查中访谈员或回应者则难以控制这类问题。
- 在问题表述或一系列提问中，可以把从先前的问题甚至先前的调查中获得的资料考虑进来。
- 如果给出的资料不一致，访谈员可以立即纠正。
- 电脑可以将资料添加到资料文档中以备立即进行分析。

也有不使用这个系统的恰当理由。其中一个可能最重要的理由是必须花时间给电脑辅助方案编程，而且程序必须准确无误。因为在进行访谈时，访谈员无法像处理印刷错误一样处理程序错误。在没有访谈员时，回应者直接输入他们的答案就有可能出现更多的错误。尽管一些简单的少有跳题的调查工具不会导致太多的程序测试问题，但是在开始访谈前仍需相当多的时间来进行程序测试和修正。

另外，整个资料输入过程中没有质量控制。除了确保输入的是合法代码并具有内在一致性之外，访谈员使用电脑辅助系统进行的任何资料输入或任何编码决策都没有经过检查。尽管出现打字错误的几率相对来说比较低，但编码决策的质量却需要特别注意(Dielman & Couper，1995)。由于对编码决策缺少控制，当提问开放式问题时，CATI 和 CAPI 的访谈者通常都要将答案逐字输入电脑以备后来进行编码。尼科尔斯(Nichols，1988)、贝克和莱弗(Baker & Lefes，1988)、萨瑞斯(Saris，1991)、卡特林和英格拉姆(Catlin & Ingram ，1988)、尼科尔斯、贝克和马丁(Nichols，Baker，& Martin，1997)，特别是还有库珀等(Couper et al. ，1998)编辑出版的专著，都对电脑辅助系统的特点、使用方法和经验做了细致的总结。

资料清理

一旦完成对访谈的编码并把资料输入磁带或者磁盘文件存储后，此时就需要对资料进行检查。最重要的检查是弄清资料文档是否完整和有序。还要检查每个方面以确保只有合法代码出现。

即使是在资料输入时就进行过检查，最好还是对整体布局进行一个全面检查，以确保各方面工作符合如期目的。当然，如果资料输入时没有进行过检查，那么对资料的内在一致性也应当进行检查。

发现错误时就要诊断其来源并进行更正（注意：在使用 CATI、CAPI 或者 CASI 系统时不可能做到这一点，因为没有保留可以确证的答案脚本）。由于在改正过程中会再次出现错误，所以还得再检查一遍。如果资料量很大，这个清理过程会很耗时并且容易产生误差。最为理想的是能够在资料输入时发现错误，这样就可减少对后续资料清理的依赖。

编码和资料归并导致的调查误差

由于编码和资料归并能在严密监督下进行并能被全面彻底的检查，这就使它有可能成为调查过程中几乎没有误差的一部分。并且，编码和资料归并的成本通常只占整个调查成本的一小部分。

当处理封闭式答案时，资料输入过程中出错的几率低于 1%。如果那些数字是直接输入并全部经过核实，抄写员在抄写过程中也进行检查，在最后得到的资料中出现误差的概率就更低。

给开放式应答编码时信度会因问题的质量、编码的质量和对编码员的培训和监督不同而有所不同。如果研究者有好的问题，并且代码类型都很清楚，那么就可以预计编码有 90%以上的信度，也就是说编码员和编码检查员不同意那种分类的比例低于 10%。没有经过严格训练和没有监督的编码员可能会出现较多的编码错误。编码依赖于已知的完整的定义，例如，职业类型、健康状况或特定的犯罪等，因此研究者要特别注意对编码员的培训和对编码工作的检查。

如果经过仔细检查，资料输入过程就有可能达到几乎没有错误。虽然有一些操作员能以相当高的精确度（出错率低于 1‰）输入数据，但是仍然不能完全认为资料输入是绝对准确的。

对编码和资料输入过程的选择通常是由其他原因决定的，而不是考虑其误差最小化。在访谈中构建文档的高速度和找到错误

的高概率，是 CATI 和 CAPI 系统吸引人的地方，这就像电脑辅助使问题陈述具体化和使提问有序化的长处一样。但是，单从减少误差这个角度来看，当调查包含了大量的编码决策时，可以选择编码员直接输入数据，然后 100%校验其工作（编码和资料输入）的两步法。没有别的系统能够提供给所有编码决策和所有资料输入一个真正独立的核查。



9 调查研究中的道德问题

Ethical Issues in Survey Research

跟所有的社会研究一样，调查研究应该在避免参与者、回应者和访谈员冒风险的设计方式中展开。本章将简要阐述开展合乎道德的调查的一些措施。

在以人为研究对象的所有调查中，研究者必须注意在研究过程中运用道德的方式。基本的方针是研究者要确保在调查结束后没有任何人受到任何不利的影响，而且一个优秀的研究者还应注意到将研究过程的积极影响最大化，这在某种程度上是可以做到的。

在美国，几乎所有的大学和其他很多机构，在进行联邦政府赞助的调查时都有一个社会公益审议会(IRB)，它负责监督以人为对象的研究。当提出研究建议时，调查主持人在开始收集资料前，必须把议案交给 IRB 进行评估。

IRB 评估是为了保护研究对象、研究者和研究机构。他们最关心的问题是研究中参与者可能会冒的各种风险。只要涉及以人为研究对象的研究活动，像教育测试、调查过程或者对公众行为的观察等要想没有风险，必须：

- 记录资料时要识别研究对象……
- 公开研究对象所提供的答案……不会使其陷于不道德的境地，或违背公民责任，或有损其经济地位、就业或声誉(健康和人类服务部，1996，P. 5)。

只有遵循这些方针，很多调查可以免除风险。但是，调查研究涉及多方面主体，因此必须要向 IRB 提供材料，以便他们(通常是委员会主席)确定调查是否符合这些标准。如果调查确实有潜在的风险，或者牵涉到易受伤害的群体，IRB 有责任评估所有的过程，以确保研究对象受到很好的保护。

本书不可能阐述所有涉及特殊群体研究中的问题。当对儿童、智力发展迟缓的人、精神病患者、罪犯或其他特殊总体进行研究时，研究者应从别处获得指导。希伯(Sieber，1992)更详细地论述了怎样以道德的方式收集资料。下面是一些对一般总体进行调查时应遵循的道德原则，这是所有调查研究者都应该熟悉的。

告知回应者

一般说来，调查研究过程需要调查对象的自愿合作。合乎道德的调查研究的一个基本前提是应该告知回应者他们自愿参加的是什么活动。具体来说，回应者在回答问题之前应该知道以下信息：

- 进行这项研究的机构名称。如果有访谈员参与，回应者也

应该知道访谈员的名字。

- 主办人,即那些支持和资助研究的人。
- 对研究目的准确而又简明的阐述。研究是否试图增进一般的或基本的知识?或者是否有些研究援助的计划或行动过程?这个研究覆盖的问题或主题是哪些?研究中设计了哪些要回答的问题?
- 对答案的保密程度。如果在资料的保密性上有些限制,研究者一定要清楚地告知回应者。
- 对参与调查研究的自愿性以及对不参与调查研究的人无消极影响的担保。
- 保证回应者能跳过任何他们不愿意回答的问题。

如果研究设计允许的话,这些资料应该提前寄出或者直接送达回应者。无论如何,还是应该要求访谈员(若有访谈员)在开始访谈前再给回应者重复一下上述几点。

最后,对于合同形式的调查或许最好要签字同意。一般说来,在完成访谈前不要叫样本调查的回应者签名。显然,从电话调查中得到签名是不可行的。然而即使是在个人访谈调查里,大多数考虑周全的评审委员会认为回应者没必要签名同意,因为在大多数情况下,调查参与者所冒的风险都很小。另外,回应者在被问到每一个新问题时,有重新决定是否参加调查的机会。

也有一些例外。在下列情况时,研究者和 IRB 的委员们很可能希望回应者以书面形式证明他们对调查有了足够的了解:

- 收集特别敏感资料的时候。如果公开这些资料,可能会使回应者感到尴尬或对他们不利。
- 资料的保密程度不够的时候。
- 在调查对象缺少判断或能力以拒绝参与的时候。例如,调查对象是儿童、罪犯、雇员、学生时。
- 在使用以前收集的非研究性资料的时候。例如,寻找医疗记录作为调查答案的补充。

上面提到的都是一些例外的情况,并非一般情况。只要保证在回应者同意参与之前能了解研究草案,大多数调查访谈不需要书面签名合同。此外,希伯(Sieber,1992)充分讨论了书面合同问题。

保护回应者

如果是从一个诸如保险计划成员或某机构雇员等的名单上抽取样本，必须坚持的一个最基本的道德原则是，不管样本成员是否同意参与调查，他们的生活在任何方面都不应该受到不利的影响。也就是，样本成员是否回应都不应该被非研究团体的人知道。如果被抽取的样本成员不参加调查，也要保证他们不会有不利的结果（诸如他们的健康福利、服务、工作情况或者级别等）。

除此以外，有关保护调查回应者最主要的一点就是对他们所提供的资料的处理方式。与使用书面问卷或访谈日程表相比，直接向电脑输入答案一般较容易使资料具有保密性。细心的调查研究者减少泄密事情发生往往会采取以下步骤：

- 所有接触资料的人或收集资料的人都必须承诺为资料的机密性负责，并写下保证书。
- 尽量减少答案和身份识别符号之间的关系。名字或地址是最普通的身份识别符号，通常书面调查都不要求写姓名。当可以避免姓名时，很多调查机构都不在研究过程中的任何部分使用回应者的姓名。
- 有详细的名字或地址等身份识别符号时，研究者会把它们填入表格使它们与真正的调查答案分开。研究者会尽快把身份识别符号（不同于被编码的身份识别数字）从完成了的调查中完全移出。
- 如果使用名字或地址鉴别样本或收集资料，如果不再需要时，研究者会马上把它们删除或销毁。
- 确保完成了的调查报告不被非项目成员轻易接近。如果是书面报告，最好把它们锁进档案柜。
- 不让那些可以从答案的特点猜测出回应者的人看到真实的调查回答（例如雇员调查里的管理人员或学生调查里的老师）。
- 真实的资料文档通常都会为每个回应者标上某种身份识别数字。身份识别数字与样本地址或回应者身份识别符号之

间的联系不要被资料文档的一般使用者获得。

- 在分析过程中，当公开一些易被确认的人的资料时研究者要特别小心。
- 当项目完成时，或当调查手段的使用结束时，研究者要确保将调查研究手段销毁或者对其继续保密存放。

显然，一些特定项目的处理方法可能有所不同，但是，上述处理方式是任何负责的调查研究项目都应当遵循的。*

对回应者的奖励

在大多数调查中，回应者获得的奖励主要是精神上的，如他们在访谈过程中得到了享受或感觉他们做了值得做的事情。有时也会提供物质上的奖励，例如，酬金、奖品，或帮扶。当向回应者提供奖赏时，应当注意要用一种不伤害对回应者答案信任的方式提供。除此以外，研究所应坚持的最重要的道德准则就是不夸大奖励并按许诺寄出奖励。通过描述调查的用途以寻求回应者合作的研究者，尤其要信守对资料进行合理的分析和传播的承诺。

对访谈员的道德责任

研究者要在两个方面对访谈员负责：第一，因为访谈员有责任把研究内容告知回应者，所以研究者的责任就是确保访谈员全面准确地了解有关研究的资料，不应该使访谈员处于一种被欺骗的、

---

* 法庭有可能向研究者索取问卷和记录，而研究人员可以以威胁到承诺的机密性为由保护问卷和记录。如果研究牵涉到特别敏感的材料，例如，司法研究中可能涉及的毒品或罪犯，研究者可以请求联邦或州机构保护问卷和记录不受传唤，或者在传票发出前，研究者可以销毁身份识别符号与答案间的联系。如果保持这个联系十分重要，例如一些纵向研究需要重复与回应者的联系，研究者要把保存有个人和答案之间的联系的文档送到国外或在美国法庭权限以外的地方。此外，希伯(Sieber,1992)的文章中提到了此问题的更多细节。

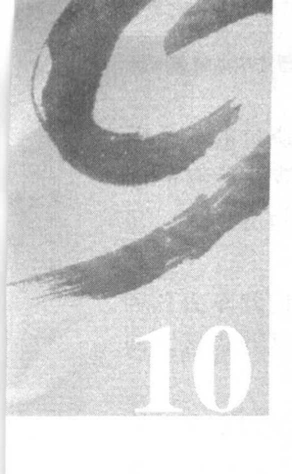
误导的或不适当的境地。第二,研究者必须解决访谈员的安全问题和对犯罪恐惧的问题。因为一般的家庭样本要涉及所有的区域,这样访谈员可能会去那些令他们感觉不安全的地方。这时研究者可以遵循以下一些原则处理:

- 在确定访谈员感觉不安全之前,理应要求他们驾车去各种样本所在地参观,因为样本所在区域不同,街区与街区之间也不同。
- 要详细明确地告诉访谈员,在他们觉得不安全的情况下不会要求他们去工作。解决方法包括:避免晚上做访问,尽量周末去访问上班族,与其他访谈员一起访谈或出钱找护送员。还可以让访谈员与现场监督者一起找出他们觉得安全的访谈方法。
- 事前应该给访谈员详细介绍避免他们成为受害人的切实可行的措施。

幸运的是,访谈员真正出事的情况是非常少有的,但是社会里的确有犯罪现象存在,因此研究者和访谈员都需要感到访谈员已经清楚了去哪些地方或做哪些事情可能不安全,并且要使访谈员没有必须去那些地方和做那些事情的压力。

总 结

调查研究中的道德问题与社会科学中的一般道德问题没有什么不同。在大多数调查中,回应者(或访谈员)的真正的风险和可能的花费都很小,但仍然需要采取一些基本的步骤来减少参与者或社会科学研究者的风险。上面列出的步骤绝不是详尽无遗的,但是,在追求最大利益避免不必要花费的同时,真诚地对待每个人的基本态度,是任何调查研究应努力做到的一个重要方面。



10 提供有关调查方法的资料

Providing Information About Survey Methods

提供所使用的影响估计值的调查措施的全部细节，是报告调查估计值的研究者的一种科学上的义务。另外，研究者应展示和报告与他们的数值的精确度和准确度有关的计算方法。本章的讨论是一个全面的调查研究方法论描述不可或缺的。

研究者做出的方法论决策可以被称为绝对错误的情况是很少出现的。对于一些研究情境来说，本书讨论到的任何折衷方案对于资料收集来说是合理和划算的。尽管不能离开具体情况指责研究设计决策，但是如果研究者不能完整地描述资料收集过程就应该受到指责，因为让调查资料的读者和使用者都能够了解全面和完整的资料收集过程是十分重要的。

一个好的调查方法描述一般有两种功能。第一种功能，是让人们了解到样本估计值怎样很好地描述了总体。只是给出作者在这个方面的结论是不够的，而应该给出与精确度和偏差有关的详细运算，以便让读者做出自己的评估。第二种功能，是提供程序上的细节，以便其他研究者可以重复相同的资料收集过程或找出可能会影响可比性的各调查间程序上的差异。

经常可以发现只对一项调查的样本规模予以报告的情况，而更尽责的研究者应对抽样策略和回应率进行说明。虽然合理的详细程度会因使用资料的不同方式而改变，但是对任何调查都应该提供下列简明提纲所概括的信息：

- 样本框（即从中抽取样本的那些人），有机会从样本框中被抽取出的那部分人的估计值，以及已知的关于被排除的人与总体之间差异的任何信息。
- 抽样过程，包括在简单随机抽样的基础上的变异，诸如整群法、分层法，或者从总体的各子群中以不同比例抽样的方法。
- 现场结果，最初指定样本的处理：回应者的总数、无回应者的总数，以及不回应的主要原因。如果由于样本框包括了不合格的单元（例如，与住家无关的电话号码）导致不能精确计算回应率，研究者应该报告不合格的单元的数目和最可能的回应率估计值。美国公共意见研究协会 2000 年出版了一本论述报告回应率的专著，以提高报告的一致性以及研究者报告研究结果时所用术语的一致性。
- 被分析的问题的准确表述：对于一个重要的报告来说，应该在附录中再现整个调查的手段。

除了对资料收集过程的事实说明以外，方法论附录还应包括

其他五类内容：

第一，由于很多报告不是提供给调查研究方法论专家看的，所以，简要概述调查中可能出现的误差种类，通常是一个调查报告的方法论部分的恰当说明。

第二，应该包括与样本独特设计有关的抽样误差量的估计值。如果样本设计用的是分层法和整群法，或者抽取中使用了不同的概率，这些设计特点对调查测量会有不同影响。研究者一般应计算这些设计对一定数目的调查测量的影响，其中包括一些他们最希望和最不希望的影响。然后他们应展现这些设计影响的项目或报告出设计影响的范围，并且概括受样本设计影响最大的项目种类。

第三，提供那些会影响到所收集资料的访谈者的信息。至少要说明以下内容：使用访谈员的数量；新培训的和经验丰富的访谈员各自的数量；访谈员接受的培训程度；是否从他们的工作中进行了抽样检查。如果访谈员的一些特征和调查内容有关，最好也提供他们的年龄、性别和种族等特征。

但是正如在第 7 章里提到的一样，如果分配给访谈员的回应者是基于方便的原则进行的，那么访谈员的影响是不能被可靠计算出来的。格罗夫斯（Groves，1989）总结了很多有关访谈员的影响的资料，这些资料表明了访谈员对估计值影响的范围。如果能通过更多的研究得到评估访谈员影响的方法，并且将最后的评估值列为调查方法论报告的一项重要内容，应该是比较理想的。

第四，研究者应该尽可能地告诉读者无回应对样本估计值的影响。另外，如果研究者对能够提供无回应者资料的人群进行了抽样，也可将得到的资料公布出来。应鼓励访谈员尽量从那些拒绝回答的人那里得到不回答的原因，以便研究者可以大概说出无回应者和回应者的区别。如果能够从其他来源得到有关总体的统计学数据，例如相关的近期普查数据，那么研究者可以将样本与这些数据进行比较以评估出无回应对样本的影响。

第五，一个好的方法论附录还应该包括一些有关调查中使用到的主要测量方法的信度和效度的资料。一般来说需要报告三种相关资料：

首先，如果问题接受了认知实验室测试或系统的检测，那么报告中应该包括对这一事实及其结果的介绍。简单报告已做的问题评估的类型对使用这个结果的人十分有益。值得报告的内容包括：问题被证明是可以理解的；前测中的行为编码表明问题是按所写的进行提问的，并且通常易于回答；有时前测表明问题有疑点，但却保留了。显然，这些信息对使用研究结果的人是非常有帮助的。

其次，研究者应提供评估答案效度所做的分析。这些答案与其他问题的答案在可预测的方式上关联的程度，显示出这些问题所测量的正是研究者所希望测量的。韦尔（Ware，1987）提供了全面评估信度和效度所必须的分析种类的一个很好的总结。

再次，虽然在调查中很少能直接评估出事实资料报告的准确度，但是通过引用查核记录得来的类似项目的资料，可以给读者提供一个根据，使他们估计出调查估计值中的误差数量和方向。

关于问题能够在多大程度上测量出所希望测量的内容，到目前为止，还缺乏这个方面的系统性研究资料。可以说，调查报告的大部分内容保证了表面上的效度——也就是说答案能够符合问题设计者的要求。以评估问题和答案为目的的资料收集和分析方法应成为调查方法论报告的常规部分。

将这里列出的所有信息包括在一个调查资料报告里，会显得过于详细，通常会承受杂志要求缩短文章的压力。然而，所有这些信息明显地关系到调查评估值可靠质量的评价。在一个全面的调查分析报告中，完整的方法论附录应是不可少的。如果发表时需要缩短篇幅，至少也应该有关于资料收集过程细节的方法论报告。

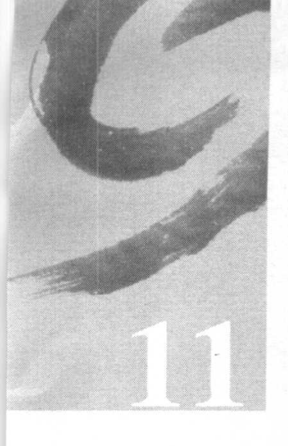
总之，应该注意，收集和分析有关测量资料的过程很重要，对它们的报告同样很重要。显然，采取措施以使误差最小化是必要的，但是完全没有误差的调查是不可能的。报告测量是如何顺利进行的，评估结果中误差的数量和类别，是保证对调查资料合理使用的一个重要方面，也是构建进一步提高调查测量水平所依赖的知识的一个重要方面。

因此，在报告研究结论时，研究者有义务在报告复现其结论所需信息的同时，也要报告评估其调查资料质量所需要的信息。要

达到后一个目的必须仔细地描述所使用的调查程序;而实现前一个目标需要特别重视误差测量和结果交流。当一个报告对诸如问题是否清晰或访谈员是否影响到答案等有关误差问题避而不谈时,大多数读者会认为这个报告是没有问题的。研究者至少能让读者知道影响调查估计值误差的各种不同来源(这是首要的)。但是,从长远来看,在报告调查结果时,将那些为提供详细的误差估计值所需要的步骤包括进去,应发展为一种惯例。

练 习

使用本章提到的标准,对建立在调查基础上的一本著作或一个研究报告的方法论部分的恰当性和完整性进行全面系统的评价。



透视调查误差

Survey Error in Perspective

调查的整体设计包括考虑一个调查的所有方面和选择一个合乎特定研究目标的严格标准。本章将讨论优秀的调查设计和实践中最常见的偏差，同时还要评估优秀调查设计和实践的成本节俭可能性以及它们对调查估计值的精确度、准确度和信度的意义。



调查整体设计的概念

调查的整体设计意味着在设计调查或评估调查资料质量时，研究者要着眼于完整的资料收集过程，而不只是简单地考虑调查的某一两个方面。样本(样本框、样本规模、样本设计和回应率)的质量、作为测量手段的问题的质量、资料收集(特别是有效的访谈员培训和监督措施的使用)的质量以及收集的形式构成了一系列紧密联系的问题和设计策略，所有这些方面都有影响最终调查结果的可能性。因此，对调查整体设计的全面正确评价应包含如下三方面具体含义：

- 在设计一个调查资料收集过程时，研究者要自觉地考虑到在调查设计过程中各方面的成本和方法严密性之间的平衡。如果调查的其他方面不能担保某个水平的投资对降低误差有效时，对调查的某一个方面进行这个水平的单向投资也不会奏效。
- 在评估资料质量时，研究者应当询问所有影响资料质量的决策是如何做出和实行的。
- 在报告调查的细节时，研究者应该报告影响资料误差水平的资料收集措施的各方面的细节。

透视误差

对本书讨论到的方法论选择的意义进行归纳以适用于各种不同的研究是难以做到的。因为严谨决策的成本在不同的研究情景中差异巨大，在不同的方法选择之间也存在巨大差异。同样，为做出省钱的决策而导致错误的可能性也有非常大的不同。调查的定制设计是指研究者应仔细评估各种设计选择，评估其成本和误差出现的概率，然后决定选用何种设计。

能使研究总体的每个成员都有按照某个确定几率被抽取的完

美抽样框几乎是不存在的，因此，谁可以成为样本不能脱离情境进行评估。尽管如此，研究者仍然有责任弄清其样本框的完整性以及漏掉了哪些人，并且不能暗示其样本估计值也适合于那些没有机会被抽样到的人。

在调查研究中采取节约成本的折衷措施的行为最可能发生在抽样领域。最糟糕的情况是，研究者试图从那些完全没有经过抽样的人群中收集资料以总结出结论。例如，从自愿填答期刊上调查问卷的那些读者那里收集资料。然而，对于一些很难获得合作的样本，通常的解决办法只能是使用非概率样本组成的更易得和更易取得配合的人群，典型地表现在大多数民意、政治和市场研究测验中。事实上，对于电话调查来说，样本替代不会节约多少钱，尽管它能让调查迅速地进行。而对于面对面的家庭访谈，在概率和非概率样本之间成本就会有许多差异。为了节约这种成本付出的主要代价是放弃有关资料误差的统计数据的信度，没有了描述样本和所由抽样总体之间的关系的科学根据。如果调查的目的是要获取关于总体的更宽泛的观点而不是获取准备获取的某些资料，这种非统计学的抽样过程可能会更好地满足这种目的要求。然而，如果兴趣在于至少是为数字定序的一些估计值，并且要考虑科学可靠性问题时，从样本替代里节约成本可能就得不偿失了。

另一个几乎类似的普遍的妥协就是接受低回应率。几乎所有关于邮件问卷早期返回的研究都表明，早期返回的问卷是一个有偏误的样本，这种偏误的存在直接影响着调查的主题。电话调查中与低回应率有关的偏误没那么明显，尽管这可能是由于它总是忽略缺乏教育的、老年的具有独特观点的人口导致的。真正的问题在于没有准确信息判断无回应何时会严重影响估计值，因此很难说回应率何时算是很低以及额外工作何时是必要的。研究者所做工作的最低限度是要保证可靠性，因此基于低回应率的调查资料很容易遭受非难，研究者不得不确定其影响有多大。

由于大多数考虑调查设计的研究者会考虑到抽样，所以这里就不花更多的时间讨论这个问题。如果要选取精密复杂的样本，就需要合格的抽样统计学家的帮助，因为只有这样的统计学家才能考虑到计算抽样误差的不同抽样策划之间的平衡。这里应该指

出的一点是，总体样本的大多数都牵涉到整群抽样。事实上，多阶整群样本是处理大范围抽样问题的最好办法。但是，经常可以发现，在一些报告中整群抽样的影响在评估抽样误差时被忽略了。如果研究者没有一个有效的简单随机样本（这种情况经常出现），那么在评估标准误差时，他们就不能使用简单随机抽样假设。

资料收集模式的选择是影响调查成本的最重要选择之一。虽然在很多年里，个人访谈调查的各种方法曾被认为是进行一般总体调查的惟一有效的方法，但是现在在大多数调查机构中使用电话调查的方法比个人访谈调查更为流行。

在很多方面，电话访谈已被证明是一种产生调查资料的有效途径。但是，电话调查的成本往往使得这一方法不是最好的选择。情况不同，电话访谈可能会存在一些不足：排除了总体中没有电话的部分（样本框的问题）；会产生较高的无回应率（虽然这并不常发生）；会使研究者收集的资料更少（因为电话访谈的时间通常比个人访谈时间短）。但总地来说，电话调查过程是很多项目的正确选择。就大多数节约成本的手段来说，有时选择电话调查的非货币价格会很高。

就某些总体的资料收集来说，邮件和互联网显然是有诱惑力的方式。由于其相对低廉的价格，邮件和互联网常被用来与各种花费高的调查方法结合使用：用廉价的邮件和互联网方式从使用邮件和互联网的人群中收集资料；那些不太愿意回答问题的人或者不便于用邮件和互联网方式回答问题的人（例如，因为他们很少接触互联网）的资料，则用其他的资料收集方式收集。优秀的调查设计要能以省钱高效并获得高回应率的方式解决资料收集方式与样本、研究主题、工具需求之间的匹配关系问题；而糟糕的调查设计则只依靠一种对研究总体的某些部分不合适或效率低的方式收集资料。

访谈员的质量也是调查研究里最易忽略的问题。事实证明，良好的培训和监督可以对访谈员进行的访谈产生的巨大作用。例如，对于大多数一般的调查问题来说，没有经过良好培训或监督的访谈员可能会增加围绕估计值误差的幅度，就像取了一个更小的样本一样。这就是说如果存在访谈员对答案的严重影响，一个

1 000人的样本产生的样本精确度可能只及一个700人的样本。虽然访谈质量的重要性随调查的内容和问题类型的不同而不同，但是在大多数一般目的的调查中，都有部分问题受到了访谈员的严重影响。在这情况下，削减对访谈员的培训和监督的费用就是一个错误的决定。

另外，访谈人员的素质也会影响到回应率。使用能够取得回应者合作的优秀访谈员，是保证高回应率的最简单的一种方法。因此需要对访谈员严格监督、再培训，并排除那些不善于取得合作的访谈员，以便减少无回应率和可能出现的额外支出。

将访谈员向回应者解释访谈任务的方式格式化为一些标准化程序，可以用于调查研究。事实证明，这种技术是一种实际上不需成本而能提高回应者表现水平的方式。

最后，对调查问题全面的评估可能是最重要和最划算的减少调查误差的一种方法。例如：通过对焦点小组和认知访谈的合理运用，研究者可以提高回应者对问题理解的一致性和答案的有效性。而在现场前测中对行为的系统研究可以鉴别出访谈员是否以一致的方式提出问题或把问题展现给回应者的。提高问题的质量既可以提高回应者报告的效度又可以减少访谈员对答案的影响。

总　结

优秀的设计和实践的目的是让每一块钱能够最大可能地获得准确性、可靠性和可重复性。而一个特定研究的精确性、有效性和可重复性是由提出的问题和资料的使用方式决定的。

人们有时会读到一些文章，它们认为社会科学是不精确的，并且还对社会科学和物理科学的测量过程做过引起人们反感的比较。尽管这种见解是常见的，但它们通常也是无知的。物理科学和生物医学中的一些基本测量，诸如血液样本中铅的含量、血压的读数、X光的读数和金属弹性的测量等，都被证明具有较低的可信度。这些领域里的测量都可能产生可好可坏的结果，这取决于使用什么样的方法论以及设计测量过程谨慎与否，这个方面特纳和

马丁(Turner & Martin ,1984)提供了很多例子。对于调查研究来说,情况也是一样。

要想将调查里的误差最小化,通常要花费较高的成本。在有些情况下,最好的测量价值却可能不抵对它的巨大支出。通过抽样家庭来抽样总体,通常会排除与家庭没有关系的人。在大多数地区,这部分与家庭无关的人只占总体的一部分,但是要设计一种方法去抽样这些与家庭无关的人,花费会是很高的。

无回应率可以减少到0。人口普查局在进行国家健康访谈调查时,回应率超过了90%,只是在一些主要的大城市回应率要低一些。而且如果有人想投入足够的时间和金钱,即使是在大城市的最不好客的街区,其回应率仍可能会接近100%。然而也有花费极高,潜在误差减少却很有限的情况。

幸运的是,回应者能够并且愿意回答社会科学家和决策者感兴趣的很多问题。但是对于另外一些问题,假如回应者能够并且愿意在标准调查过程中回答这些问题,就会给研究者带来很大的方便,可是回应者往往却不这样做。例如,使用标准调查技术调查酒后驾车和破产时,这些情况被回应者明显少报了(Locander, Sudman, & Bradburn, 1976)。或许一项研究可以拥有某种方法,以引导回应者准确地报告这些事实,但这样需要研究者付出比通常更大的努力来赢得回应者的合作。研究者还得依据将来使用资料的方式对获得多大的准确度和详细度做出决策。读者可以参看格罗夫斯(Groves, 1989)的文章,其中对调查成本和调查误差之间的关系做了更为广泛的分析。

据说调查研究的局限性在于人们在调查情景中能够并且愿意告诉研究者一些什么。然而,这些局限性是被夸大了的。在用标准化调查措施能够精确地测量哪些内容上,的确存在着一些局限性,但是这些局限性可能更多地与预算和研究者在测量过程中所付出的努力有关,而不在于标准化调查是否实际可行。

研究者对设计评估的前提,是要看所做的折衷选择是否正确、是否明智、是否会形成符合预期目的的资料。值得指出的是,调查误差并不仅仅来源于决策的严密性和成本节约方面,还来源于对无回应率影响的评价、问题设计以及访谈员行为对调查估计值质

量的影响。由于全然缺乏对资料收集这些方面的重视，就经常会得到缺乏可靠性或不符合信度的最新标准的调查资料（例如参看 Bailar & Lanphier，1978；Turner & Martin，1984）。

对本书阐述的设计观念的正确评价应包括以下方面：

- 在资料收集的特点中，没有哪一点过于粗糙或者薄弱，以至使得研究者无力使用资料达到调查目的。
- 对资料收集的各个方面的设计要保持相对的一致，以便避免出现对某一方面的投入提高了精确度却无法被其他方面精确度变化证明是合理的情况。
- 调查资料的使用者能慎重地使用样本调查的估计值，能考虑到调查误差的可能来源，并能认识到估计值精确度和置信度上的局限性。

最后，希望研究的使用者对自己提出的关于任何调查中资料收集的那些问题有一个全面的把握；希望研究者对设计决策细节的重要性有更好的把握；希望所有的读者可以用新学到的知识致力于更优秀的调查设计和实施。

参考文献
References

American Association for Public Opinion Research. (2000). *Standard definitions: Final dispositions of case codes and outcome rates for surveys*. Ann Arbor, MI: AAPOR.

Anderson, B., Silver, B., & Abramson, P. (1988). The effects of race of the interviewer on measures of electoral participation by blacks. *Public Opinion Quarterly*, 52(1), 53-83.

Aquilino, W. S. (1994). Interview mode effects in drug surveys. *Public Opinion Quarterly*, 58(2), 210-240.

Aquilino, W. S. (1998). Effects of interview mode on measuring depression in younger adults. *Journal of Official Statistics*, 14(1), 15-30.

Aquilino, W. S., & Losciuto, L. A. (1990). Effects of mode of interview on self-reported drug use. *Public Opinion Quarterly*, 54(3), 362-391.

Bailar, B., & Lanphier, C. (1978). *Development of survey research methods to assess survey practices*. Washington, DC: American Statistical Association.

Baker, R. P., & Lefes, W. L. (1988). The design of CATI systems: A review of current practice. In R. M. Groves, P. N. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nichols II, & J. Waksberg (Eds.), *Telephone survey methodology*. New York: John Wiley.

Beatty, P. (1995). Understanding the standardized/non-standardized interviewing controversy. *Journal of Official Statistics*, 11 (2), 147-160.

Belson, W. A. (1981). *The design and understanding of survey questions*. Aldershot, England: Gower.

Berry, S., & Kanouse, D. (1987). Physician response to a mailed survey: An experiment in timing of payment. *Public Opinion Quarterly*, 51(1), 102-114.

Billiet, J., & Loosveldt, G. (1988). Interviewer training and quality of responses. *Public Opinion Quarterly*, 52(2), 190-211.

Bishop, G. F., Hippler, H. J., Schwartz, N., & Strack, F. (1988). A comparison of response effects in self-administered and telephone surveys. In R. M. Groves, P. N. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nichols II, & J, Waksberg

(Eds.), *Telephone survey methodology* (pp. 321-340). New York: John Wiley.

Blom, E., & Lyberg, L. (1998). Scanning and optical character recognition in survey organizations. In M. Couper et al. (Eds.), *Computer-assisted information collection* (pp. 449-520). New York: John Wiley.

Bradburn, N. M., & Sudman, S. (1992). The current status of questionnaire design. In P. N. Beimer, R. M. Groves, L. E. Lyberg, N. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 29-40). New York: John Wiley.

Bradburn, N. M., Sudman, S., & Associates. (1979). *Improving interview method and questionnaire design*. San Francisco: Jossey-Bass.

Brick, J. M., Waksberg, J., Kulp, D., & Starer, A. (1995). Bias in list-assisted telephone samples. *Public Opinion Quarterly*, 59, 218-235.

Bryson, M. (1976, November). The Literary Digest poll: Making of a statistical myth. *American Statistician*, 184-185.

Burton, S., & Blair, E. (1991). Task conditions, response formulation processes, and response accuracy for behavioral frequency questions in surveys. *Public Opinion Quarterly*, 55, 50-79.

Cannell, C., & Fowler, F. (1964). A note on interviewer effect in self-enumerative procedures. *American Sociological Review*, 29, 276.

Cannell, C., Groves, R., Magilavy, L., Mathiowetz, N., & Miller, P. (1987). An experimental comparison of telephone and personal health interview surveys. *Vital & Health Statistics*(Series 2, No. 106). Washington, DC: Government Printing Office.

Cannell, C., Marquis, K., & Laurent, A. (1977). A summary of studies. *Vital & Health Statistics* (Series 2, No. 69). Washington, DC: Government Printing Office.

Cannell, C., & Oksenberg, L. (1988). Observation of behaviour in telephone interviewers. In R. M. Groves, P. N. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nichols II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 475-495). New York: John Wiley.

Cannell, C., Oksenberg, L., & Converse, J. (1977). *Experiments in interviewing techniques: Field experiments in health reporting*, 1971-1977. Hyattsville, MD: National Center for Health Services Research.

Catania, J. A., Gibson, D., Chitwood, D., & Coates, T. (1990). Methodological problems in AIDS behavioral research: Influences on measurement error and participation bias in studies of sexual behavior. *Psychological Bulletin*, 108(3), 339-362.

Catlin, G., & Ingram, S. (1988). The effects of CATI on costs and data quality. In R. M. Groves, P. N. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nichols II, & J. Waksberg (Eds.), *Telephone survey methodology*. New York: John Wiley.

Conrad, E G., & Schober, M. E (2000). Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly*, 64(1), 1-28.

Converse, J. (1987). *Survey research in the United States*. Berkeley: University of California Press.

Converse, J., & Presser, S. (1986). *Survey questions*. Beverly Hills, CA: Sage.

Couper, M. P. (1999). The application of cognitive science through computer assisted

interviewing. In M. G. Sirken et al. (Eds.), *Cognition in survey research* (pp. 277-300). New York: John Wiley.

Couper, M. P., Barkin, R. P., Bethlehem, J., Clark, C. E, Martin, J., Nicholls, W. L., & O'Reilly, J. M. (Eds.). (1998). *Computer assisted survey information collection*. New York: John Wiley.

Couper, M. P., Hansen, S. E., & Sadowsky, S. A. (1997). Evaluating interviewer use of CAPI technology. In L. E. Lyberg, P. Beimer, M. Collins, et al. (Eds.), *Survey measurement and process quality* (pp. 267-286). New York: John Wiley.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychiatrika*, 16, 297-334.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

de Leeuw, E. D., & van der Zouwen, J. (1988). Data quality in telephone and face to face surveys: A comparative meta-analysis. In R. M. Groves, P. N. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nichols II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 283-299). New York: John Wiley.

DeMaio, T. J., & Rothgeb, J. M. (1996). Cognitive interviewing techniques: In the lab and in the field. In N. A. Schwarz & S. Sudman (Eds.), *Answering questions* (pp. 177-196). San Francisco: Jossey-Bass.

Department of Health and Human Services. (1996). *Protection of human subjects* (Title 45, Code of Regulations, Part 46). Washington, DC: Government Printing Office.

DeVellis, R. E (1991). *Scale development: Theory and applications*. Newbury Park, CA: Sage.

Dielman, L., & Couper, M. P. (1995). Data quality in CAPI survey: Keying errors. *Journal of Official Statistics*, 11(2), 141-146.

Dillman, D. A. (2000). *Mail and Intemet surveys: The tailored design method*. New York: John Wiley.

Dillman, D. A., Carpenter, E., Christensen, J., & Brooks, R. (1974). Increasing mail questionnaire response: A four state comparison. *American Sociological Review*, 39(5), 744-756.

Dillman, D. A., & Miller, K. J. (1998). Response rates, data and feasibility for optical scannable mail surveys for small research centers. In M. P. Couper et al. (Eds.), *Computer assisted information collection* (pp. 475-498). New York: John Wiley.

Dillman, D. A., & Tarnai, J. (1991). Mode effects of cognitively designed recall questions: A comparison of answers to telephone and mail surveys. In P. N. Beimer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 367-393). New York: John Wiley.

Dykema, J., Lepkowski, J. M., & Blixt, S. (1997). The effect of interviewer and respondent behavior on data quality: Analysis of interaction coding in a validation study. In L. Lyberg, P. Beimer, M. Collins, et al. (Eds.), *Survey measurement and process quality* (pp. 287-310). New York: John Wiley.

Edwards, W. S., et al. (I 994). Evaluation of National Health Interview Survey diag-

nostic reporting. In *Vital & Health Statistics* (Series 2, No. 120). Hyattsville, MD: National Center for Health Statistics.

Edwards, W. S., Winn, D. M., & Collins, J. G. (1996). Evaluation of 2-week doctor visit reporting in the National Health Interview Survey. In *Vital & Health Statistics* (Series 2, No. 122). Hyattsville, MD: National Center for Health Statistics.

Erlich, J., & Riesman, D. (1961). Age and authority in the interview. *Public Opinion Quarterly*, 24, 99-114.

Filion, F. (1975). Estimating bias due to nonresponse in mail surveys. *Public Opinion Quarterly*, 39(4), 482-492.

Forsyth, B. H., & Lessler, J. T. (1992). Cognitive laboratory methods: A taxonomy. In P. N. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 393-418). New York: John Wiley.

Fowler, F. J. (1991). Reducing interviewer related error through interviewer training, supervision, and other means. In P. N. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 259-278). New York: John Wiley.

Fowler, F. J. (1992). How unclear terms affect survey data. *Public Opinion Quarterly*, 56(2), 218-231.

Fowler, F. J. (1995). *Improving survey questions*. Thousand Oaks, CA: Sage.

Fowler, F. J. (1998). Nonresponse bias in mail surveys of health plan members. In 1998 *proceedings, Section on Survey Research Methods, American Statistical Association*, pp. 576-580.

Fowler, F. J., & Cannell, C. F. (1996). Using behavioral coding to identify cognitive problems with survey questions. In N. A. Schwarz & S. Sudman (Eds.), *Answering questions* (pp. 15-36). San Francisco: Jossey-Bass.

Fowler, F. J., & Mangione, T. W. (1990). *Standardized survey interviewing: Minimizing interviewer related error*. Newbury Park, CA: Sage.

Fowler, F. J., Roman, A. M., & DJ, Z. X. (1998). Mode effects in a survey of Medicare prostate surgery patients. *Public Opinion Quarterly*, 62(1), 29-46.

Fox, J. A., & Tracy, P. E. (1986). *Randomized response: A method for sensitive surveys*. Beverly Hills, CA: Sage.

Fox, R., Crask, M., & Kim, J. (1988). Mail survey response rate: A meta-analysis of selected techniques for increasing response. *Public Opinion Quarterly*, 52(4), 467-491.

Friedman, P. A. (1942). A second experiment in interviewer bias. *Sociometry*, 15, 378-381.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: John Wiley.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: John Wiley.

Groves, R. M., & Kahn, R. L. (1979). *Surveys by telephone*. New York: Academic Press.

Groves, R. M., & Lyberg, L. (1988). An overview of nonresponse issues in telephone

surveys. In R. M. Groves, P. N. Biemer, L. E. Lyberg, J. T Massey, W. L. Nichols II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 191-212). New York: John Wiley.

Heberlein, T., & Baumgartner, R. (1978). Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature. *American Sociological Review*, 43, 447-462.

Henry, G. T. (1990). *Practical sampling*. Newbury Park, CA: Sage.

Henson, R., Roth, A., & Cannell, C. (1977). Personal versus telephone interviews: The effect of telephone reinterviews on reporting of psychiatric symptomatology. In C. Cannell, L. Oksenberg, & J. Converse (Eds.), *Experiments in interviewing techniques: Field experiments in health reporting*, 1971-1977. Hyattsville, MD: National Center for Health Services Research.

Hochstim, J. (1967, September). A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association*, 62, 976-989.

Hyman, H., Feldman, J., & Stember, C. (1954). *Interviewing in social research*. Chicago: University of Chicago Press.

Jabine, T B., Straf, M. L., Tanur, J. M., & Tourangeau, R. (Eds.). (1984). *Cognitive aspects of survey methodology: Building a bridge between disciplines*. Washington, DC: National Academy Press.

James, J., & Bolstein, R. (1990). The effect of monetary incentives and follow-up mailings on the response rate and the response quality in mail surveys. *Public Opinion Quarterly*, 54(3), 346-361.

Jobber, D. (1984). Response bias in mail surveys: Further evidence. *Psychological Reports*, 54, 981-984.

Kalton, G. (1983). *Introduction to survey sampling*. Beverly Hills, CA: Sage.

Keeter, S., Miller, C., Kohut, A., Groves, R. M., & Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, 64(2), 125-148.

Kish, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, 44, 380-387.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.

Kish, L. (1965). *Survey sampling*. New York: John Wiley.

Lepkowski, J. M. (1988). Telephone sampling methods in the United States. In R. M. Groves, P. N. Biemer, L. E. Lyberg, J. T Massey, W. L. Nichols II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 73-98). New York: John Wiley.

Lessler, J. T., & Forsyth, B. H. (1996). A coding system for appraising questionnaires. In N. A. Schwartz & S. Sudman (Eds.), *Answering questions* (pp. 259-292). San Francisco: Jossey-Bass.

Lessler, J., & Tourangeau, R. (1989, May). Questionnaire design in the cognitive research laboratory. *Vital & Health Statistics* (Series 6, No. 1). Washington, DC: Government Printing Office.

Likert, R. (1932). A technique for measurement of attitudes. *Archives of Psychology*, 140.

Little, R., & Rubin, D. (1987). *Statistical analysis with missing data*. New York: John Wiley.

Locander, W., Sudman, S., & Bradburn, N. (1976). An investigation of interview method, threat and response distortion. *Journal of the American Statistical Association*, 71, 269-275.

Lohr, S. L. (1998). *Sampling design and analysis*. New York: Brooks-Cole.

Mangione, T., Hingson, R., & Barter, J. (1982). Collecting sensitive data: A comparison of three survey strategies. *Sociological Methods and Research*, 10(3), 337-346.

Marin, G., & Marin, B. V. (1991). *Research with Hispanic populations*. Newbury Park, CA: Sage.

Marquis, K., Cannell, C., & Laurent, A. (1972). Reporting health events in household interviews: Effects of reinforcement, question length, and reinterviews. *Vital & Health Statistics* (Series 2, No. 45). Washington, DC: Government Printing Office.

McDowell, I., & Newell, C. (1996). *Measuring health: A guide to rating scales and questionnaires*. Oxford: Oxford University Press.

McHomey, C. A., Kosinski, M., & Ware, J. E. (1994). Comparisons of the cost and quality of norms for the SF-36 Health Survey by mail versus phone interviews: Results from a national survey. *Medical Care*, 32(6), 551-567.

Miller, P. V., & Cannell, C. F. (1977). Communicating measurement objectives in the interview. In P. M. Hirsch, F. G. Kline, & P. V. Miller (Eds.), *Strategies for communication research*. Beverly Hills, CA: Sage.

Nichols, W. L., II. (1988). Computer-assisted telephone interviewing: A general introduction. In R. M. Groves, P. N. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nichols II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 337-385). New York: John Wiley.

Nichols, W. L., Baker, R. P., & Martin, J. (1997). The effect of new data collection technologies on survey data quality. In L. Lyberg et al. (Eds.), *Survey measurement and process quality* (pp. 221-248). New York: John Wiley.

Oksenberg, L., Cannell, C., & Kalton, G. (1991). New strategies of pretesting survey questions. *Journal of Official Statistics*, 7(3), 349-366.

Payne, S. (1951). *The art of asking questions*. Princeton, NJ: Princeton University Press.

Penne, M. A., Lessler, J. T., Beiler, G., & Caspar, R. (1998). Effects of experimental and audio computer-assisted self-interviewing (ACASI) on reported drug use in the HSDA. In 1998 *proceedings, Section on Survey Research Methods, American Statistical Association*, pp. 744-749.

Robinson, D., & Rhode, S. (1946). Two experiments with an anti-Semitism poll. *Journal of Abnormal Psychology*, 41, 136-144.

Robinson, J. P. (1989). Poll review: Survey organization differences in estimating public participation in the arts. *Public Opinion Quarterly*, 53(3), 397-414.

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (Eds.). (1997). *Measures of personality and social psychological attitudes* (Vol. 1). New York: Harcourt Brace.

Royston, P. N. (1989). Using intensive interviews to evaluate questions. In F. Fowler (Ed.), *Conference proceedings: Health survey research methods* (pp. 3-8). Washington, DC: National Center for Health Services Research.

Saris, W. (1991). *Computer-assisted interviewing*. Newbury Park, CA: Sage.

Schaeffer, N. C. (1992). Interview: Conversation with a purpose or conversation? In P. N. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 367-393). New York: John Wiley.

Schober, M. F., & Conrad, F. G. (1997). Conversational interviewing. *Public Opinion Quarterly*, 61(4), 576-602.

Schuman, H., & Converse, J. (1971). The effects of black and white interviewers on black responses in 1968. *Public Opinion Quarterly*, 35, 44-68.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. New York: Academic Press.

Schwartz, N. A., & Sudman, S. (Eds.). (1996). *Answering questions*. San Francisco: Jossey-Bass.

Sieber, J. (1992). *Planning ethically responsible research: Developing an effective protocol*. Newbury Park, CA: Sage.

Singer, E., Van Hoewyk, J., Gebler, N., Raghunnathan, T, & McGanagle, K. (1999). The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics*, 15(2), 217-230.

Singer, E., Van Hoewyk, J., & Maher, M. P. (2000). Experiments with incentives in telephone surveys. *Public Opinion Quarterly*, 64(2), 171-188.

Sirken, M. G., et al. (Eds.). (1999). *Cognition in survey research*. New York: John Wiley.

Suchman, L., & Jordan, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association*, 85, 232-241.

Sudman, S. (1967). *Reducing the cost of surveys*. Chicago: Aldine.

Sudman, S. (1976). *Applied sampling*. New York: Academic Press.

Sudman, S., & Bradburn, N. (1982). *Asking questions*. San Francisco: Jossey-Bass.

Tanur, J. M. (1991). *Questions about questions*. New York: Russell Sage.

Thornberry, O. T, & Massey, J. T. (1988). Trends in United States telephone coverage across time and subgroups. In R. M. Groves, P. N. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nichols II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 25-50). New York: John Wiley.

Thnrstone, L., &Chave, E. (1929). *The measurement of attitude*. Chicago: University of Chicago Press.

Tourangeau, R., & Smith, T. W. (1998). Collecting sensitive data with different modes of data collection. In M. P. Couper et al. (Eds.), *Computer assisted survey information collection* (pp. 431-453). New York: John Wiley.

Turner, C. F., Forsyth, B. H., & O'Reilly, J. M., et al. (1998). Automated self-interviewing and the survey measurement of sensitive behaviors. In M. P. Couper

et al. (Eds.), *Computer assisted survey information collection* (pp. 455-473). New York: John Wiley.

Turner, C. F., & Martin, E. (1984). *Surveying subjective phenomena*. New York: Russell Sage.

Waksberg, J. (1978). Sampling methods for random-digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

Ware, J. (1987). Standards for validating health measures: Definition and content. *Journal of Chronic Diseases*, 40, 473-480.

Warriner, K., Goyder, J., Jersten, H., Hohner, P., & McSpurren, K. (1996). Cash versus lotteries versus charities in mail surveys. *Public Opinion Quarterly*, 60 (4), 542-562.

Weiss, C. (1968). Validity of welfare mothers' interview responses. *Public Opinion Quarterly*, 32(4), 622-633.

Willis, G. B., DeMaio, T., & Harris-Kojetin, B. (1999). Is the bandwagon headed to the methodicalogical promised land? Evaluating the validity of cognitive interviewing techniques. In M. G. Sirken et al. (Eds.), *Cognition in survey research* (pp. 133-154). New York: John Wiley.